# A Comparison Between Factor Analysis and Item Response Theory Modeling in Scale Analysis

ISBN (print) 978-90-367-7092-7
ISBN (electronic) 978-90-367-7091-0

Cover art by Hanneke Kappenburg

This research was supported by the Netherlands Organisation for Scientific Research (NWO), file 400-04-137.

rijksuniversiteit
groningen

# A Comparison Between Factor Analysis and Item Response Theory Modeling in Scale Analysis

**Proefschrift**

ter verkrijging van de graad van doctor aan de
Rijksuniversiteit Groningen
op gezag van de
rector magnificus prof. dr. E. Sterken
en volgens besluit van het College voor Promoties.

De openbare verdediging zal plaatsvinden op

maandag 23 juni 2014 om 11.00 uur

door

**Janke Carolien ten Holt**

geboren op 16 april 1982
te Groningen.

**Promotor**
Prof. dr. T.A.B. Snijders

**Copromotores**
Dr. A. Boomsma
Dr. M.A.J. van Duijn

**Beoordelingscommissie**
Prof. dr. A. Maydeu-Olivares
Prof. dr. R.R. Meijer
Prof. dr. Y. Rosseel

# Contents

# Introduction

The construction of scales is almost a routine practice in the behavioral and social sciences as well as in other disciplines, such as marketing and economics. Scales are constructed to measure properties of respondents like opinions, attitudes, abilities, and other individual characteristics that cannot be directly observed. As these properties are, in a sense, hidden, they are referred to as latent variables (LVs)[1].

Once a scale has been assembled by composing a number of items meant to be indicative of the LV, and responses to these items have been collected, the scale analyst usually employs either a factor analysis (FA) model or an item response theory (IRT) model to decide on the adequacy of the scale's measurement properties. The choice between these two types of models is generally not motivated, and seems to depend on the scientific discipline and the researcher's background.

Although clearly used for the same goals, IRT and FA models differ in their underlying assumptions and model parameters. Many a time neither sample item distributions nor model assumptions are examined to motivate a scaling model choice. Linear FA seems to be the most widely known scaling model, and is often applied to ordered categorical data, even though formally it requires the item variables to be continuous rather than discrete. Even with the results from many robustness studies on FA and IRT (e.g., Boomsma, 1983; Boulet, 1996; DeMars, 2010; Dolan, 1994; Flora & Curran, 2004; Forero & Maydeu-Olivares, 2009; Jöreskog & Moustaki, 2001; Knol & Berger, 1991; B. O. Muthén & Kaplan, 1985; Stone, 1992), open questions remain to what extent the results and conclusions of the analyses depend on the scaling method and on the distributional properties of the empirical data. Many applied researchers are not aware of the problems that can arise as a result of applying a model whose assumptions are violated. However, when employing a scale analysis, one ought to ask oneself: *Which is the best model to use, given the properties of the data?*

In this dissertation, the differences and similarities between FA and IRT, as applied to ordered categorical data, are investigated with respect to the stability and sensitivity — or robustness — of their estimation results to violations of distributional assumptions. More specifically, we focus on the distribution of the LV and the scaling items, and assess the strengths and weaknesses of the models in case of normal and

[1] The acronyms used throughout this dissertation are listed in Appendix A.

nonnormal distributions, under various conditions of scale strength and sample size. As the robustness of scaling models against violations of distributional assumptions is of practical importance to the empirical researcher, who will need to decide which model to choose given the sample data, we use the results of this study regarding the comparison of FA and IRT approaches to deduce a set of guidelines for applied scale analysis.

To thoroughly investigate and tackle the robustness questions, we evaluate FA and IRT (a) theoretically by comparing the model formulas and taking into account the most commonly applied estimation methods for the respective models, (b) empirically by reviewing journal articles on scaling research to show what is actually done in practice, (c) in model estimation results by means of a Monte Carlo simulation study to compare the robustness of either approach against violations of distributional assumptions, and (d) in practice by applying the models to empirical data.

We compare two FA and two IRT models: FA of the sample covariance matrix by means of maximum likelihood (FA-lin-ML), FA of the estimated polychoric correlation matrix by means of mean-and-variance adjusted weighted least squares (FA-poly-WLSMV), the graded response model by means of robust maximum likelihood (IRT-grm-MLR), and the nonparametric Mokken IRT model (IRT-mok). FA-lin is included in the comparison, because it is by far the standard practice in scale analysis, although it is theoretically not appropriate for the analysis of ordered categorical data. The mathematically equivalent FA-poly and IRT-grm models are included as being suitable alternatives to FA-lin for the analysis of ordered categorical data. The nonparametric IRT-mok model is included, because it is appropriate for scale analysis of all types of dominance items, not requiring specific distributional assumptions of either item or scale distributions. Moreover, comparisons between parametric and nonparametric scaling models are mostly lacking. All scaling models are introduced and compared theoretically in Chapter 1.

Once the theoretical similarities and differences have been clarified, we examine how the models are employed by applied researchers. We are interested in the frequency and motivation of choosing FA and IRT models. In addition, we assess how the analyses are performed and reported. To answer these questions, we conduct a review of journal articles in which FA and IRT models are applied in scale construction and evaluation, presented in Chapter 2.

Having demonstrated the empirical practice of applied scaling research, we turn to the field of Monte Carlo simulation research. Contrary to an applied setting, a Monte Carlo setup has the advantage of exact control over the input parameters, i.e., the population model is known. When these input parameters are compared with output parameters from the models under investigation, conclusions can be drawn on the comparative performance of the model estimators. In Chapter 3 an extensive overview is presented of previous simulation research focused on FA of ordered categorical data and two-parameter IRT models. Based on this literature review, we design our Monte Carlo simulation study, reflecting the conditions in which we expect the scaling models to exhibit suboptimal behavior.

In Chapter 4 the setup of the simulation study is presented, explicating the data generation process, and discussing the performance variables and criteria applied for the evaluation of results. We focus on the estimation of model parameters, corresponding standard errors, model fit, and LV scores, because these are all of importance to the applied researcher. A full list of expectations regarding the simulation results is presented in a final section.

In Chapters 5 and 6 the results of the Monte Carlo study are presented, identifying the quality of model performance under various conditions ranging from "ideal" (true model) to violating assumptions of LV and item distributions. The expectations given in Chapter 4 are all addressed, resulting in a comparative overview of the four models under investigation with regard to the performance variables. Therefore, in these chapters the answers to the central distributional robustness questions of this dissertation are presented most directly. From the discussion of the results a comprehensive set of guidelines is derived, included at the end of both chapters.

Having formulated the answers to the research questions posed, and having developed some guidelines, we apply these principles in a number of applied settings. To this end, we return to the practice of scale analysis in Chapter 7, demonstrating the usefulness of our Monte Carlo findings when applying FA and IRT modeling to three samples of empirical data. This proof of the pudding also serves to refine the guidelines presented earlier.

Finally, in Chapter 8, we summarize our findings, discuss the practical recommendations for employing scale analysis, and reflect on the implications of our study.

# Chapter 1

# Two Approaches to Scale Analysis

Factor analysis (FA) and item response theory (IRT) represent two groups of models that are commonly used for the construction and evaluation of scales in the social and behavioral sciences. In this dissertation, a comparison is made between two FA and two IRT models. We attend to the theoretical similarities and differences, as well as those in practice. The latter is done by means of a simulation study and by applying the models to empirical data.

In this chapter, the two approaches to scale analysis are introduced. For each approach, we start with a general introduction, after which we zoom in on the particular FA and IRT models studied in the remaining chapters. Subsequently, we provide a first, theoretical, comparison of FA and IRT. The chapter concludes with a summary and a brief outlook to the next chapters.

Since latent variables (LVs) play a central part in both FA and IRT, the first section of this chapter is dedicated to a short introduction of the concept of LVs.

## 1.1 Latent Variables

Latent variables (LVs) can be defined in many different ways, depending on the context in which they are needed. In general, an LV is an unobserved property of an object under study. What follows is an overview of the various definitions and uses of LVs.

First, LVs can be defined as entities that cannot be measured directly. Bollen (2002) argues that this definition is problematic, because it implies that researchers must be certain that, in the future, it will remain impossible to measure such a variable directly. As such an assumption presupposes knowledge about the future, it is undesirable. We can never be sure that we will remain unable to directly measure

a construct such as self-esteem, for example. However, it can be useful to treat a variable as latent *for the time being*.

Second, LVs are often referred to as hypothetical constructs (Bollen, 2002; Skrondal & Rabe-Hesketh, 2004), i.e., constructs that are conjured by researchers and do not necessarily exist. They are, for example, used in data-reduction techniques, such as principal component analysis, where an LV is defined *ad hoc* as a linear combination of observed variables. Consequently, such an LV or hypothetical construct does not need to represent any real-world phenomenon.

Additionaly, Bollen (2002) describes a number of more formal definitions for LVs, two of which will be discussed. First, the expected-value definition states that an LV is the expected value of an observable variable, which is called the true score (Skrondal & Rabe-Hesketh, 2004, p. 2). The expected-value definition is central to classical test theory. Second, the local-independence definition describes an LV as a variable that explains all the covariances between a set of items. So, conditionally on the LV(s), the items are independent. In this definition, the observed variables do not influence each other, they are only affected by the LV directly, and do not affect the LV (Bollen, 2002).

Apart from being used as the central concept of interest, LVs are included in statistical analyses for other purposes. First, LVs are employed to account for unobserved heterogeneity (Skrondal & Rabe-Hesketh, 2004, p. 9ff.). When not all covariates are taken into account in a model, some covariation between variables will not be accounted for. This covariation is then attributed to a latent residual or error term. This latent error is not a clear, singular concept, but rather a composite of various influences not included in the model. Such LVs are also referred to as random effects. Second, LVs are used to represent continuous unobserved variables that underly dichotomous or polytomous response variables. The observed variable is then perceived as a dichotomization or polytomization of an LV. Such LVs are also called latent continuous item variables. This concept is, for example, basic to tetrachoric and polychoric correlations.

Within the various definitions, there is a distinction between formative and reflective LVs (Bollen & Lennox, 1991). Formative LVs are formed by their indicators. These indicators are called cause indicators (Bollen, 1989, p. 222), since they have a direct effect on the LV. The hypothetical-construct definition and expected-value definition concern formative LVs. Reflective LVs are reflected by their indicators, called effect indicators by Bollen. Such LVs are thought to exist independent of measurement and to affect the indicator variables. The local-independence definition signifies a reflective LV.

Most behavioral and social scientists use an implicit definition of an LV, which is closest to the first definition given: Something that (presently) cannot be measured directly, exists independently of any measurement, but for which empirical indicators do exist. This definition of an LV will generally be used throughout this text. The empirical indicators or observed item variables will be taken to be reflective of an

LV. In Section 1.2.2, the concept of latent continuous item variables will be used, and denoted as such.

## 1.2 Factor Analysis

A traditional method of scale construction and evaluation is common FA as developed by Spearman (1904), later followed by (principal) component analysis (CA; Hotelling, 1933) and multiple factor analysis (Thurstone, 1947). The focal point of FA methods is the covariance (correlation) matrix between item responses, supposed to be measured on an interval scale, which is used to determine whether one or more underlying factors, or LVs, can account for the item responses.

The use of FA and CA as exploratory techniques for scale development is widespread practice in a broad area of research. The exploratory approach, however, is somewhat paradoxical, because most of the time researchers construct questionnaires from prior knowledge: The LVs to be measured are usually rather well defined. Nonetheless, the researchers' substantive knowledge is not always explicitly used in the analyses, except perhaps for the number of factors to be recovered. This practice neglects the existence and the strength of confirmatory FA techniques advanced by Jöreskog (1969). In a confirmatory analysis, one seeks to determine whether specific items, constructed on theoretical grounds, are associated according to some factor model, i.e., whether the LVs can be regarded as common factors accounting for the correlations or covariances between a set of item responses. Almost always, linear factor models are used for such an analysis, although nonlinear factor modeling procedures are available as well.

In common FA, an observed variable is decomposed into a common and a unique part. The common part is what the observed variable has in common with other observed variables. The unique part is not (linearly) related to the other observed variables, and consists of a specific part and measurement error. Generally, within the unique factor one cannot distinguish between the specific and the measurement error part.

In CA, an LV (or component) is *defined as* a linear combination of a number of observed variables. Unique parts of variables and measurement error are not modeled or accounted for. These types of models are thus more parsimonious than the common FA model. However, when a scale developer has an LV as a starting point and constructs items to make that LV observable, the common FA definition of an LV is more appropriate than the definition in CA.

Although in scale construction and evaluation many FA models are used, the research described in this dissertation is focused on *confirmatory common FA* in comparison to IRT. As items are constructed to reflect an LV that is of interest to the researcher, it is preferable to apply *confirmatory* methods to evaluate the statistical properties of the scale. *Common* FA assumes the existence of underlying LVs and is thus useful for scale analysis. Moreover, as common FA, like IRT, takes the error part of the total observed item variance into account, its use in scale construction

corresponds more to IRT than does CA, where unique parts of item variables are not included in the model.

### 1.2.1 Definitions and Assumptions

Unidimensional common FA is a model for representing scores of respondents on tests consisting of multiple items. This model assumes that the association between the item scores can be explained by one common factor, or LV, $\theta$[1] on which individual $s$ has a value (score) $\theta_s$. The strength of relation between item $i$ and the underlying LV is represented by the loading parameter $\lambda_i$. The item response is modeled to include measurement error $\epsilon_{is}$.

The model — without inclusion of intercept terms — is defined as

$$X_{is} = \lambda_i \theta_s + \epsilon_{is}, \tag{1.1}$$

where $X_{is}$ is the observed score of respondent $s$ on item $i$, $\lambda_i$ is the factor loading of item $i$ on the LV, $\theta_s$ is the LV score for respondent $s$, and $\epsilon_{is}$ is the error term. In matrix notation, Equation 1.1 is written as

$$\mathbf{X} = \boldsymbol{\lambda}'\boldsymbol{\theta} + \mathbf{E}. \tag{1.2}$$

For $I$ items and $n$ respondents, $\mathbf{X}$ $(I \times n)$ is a data matrix of item responses, $\boldsymbol{\lambda}$ is a vector of loadings of length $I$, $\boldsymbol{\theta}$ is a vector of LV scores of length $n$, and $\mathbf{E}$ $(I \times n)$ is an error matrix.

The model as given in Equation 1.2 is easily extended to include multiple, say $Q$, LVs,

$$\mathbf{X} = \boldsymbol{\Lambda}\boldsymbol{\Theta} + \mathbf{E}, \tag{1.3}$$

with $\boldsymbol{\Lambda}$ $(I \times Q)$ now a loading matrix and $\boldsymbol{\Theta}$ $(Q \times n)$ an LV score matrix. Common FA is based on the assumption that the expected value of the product of the unique part of an arbitrary item $i$ and a second item $j$ is zero (Mulaik, 2010, p. 6),

$$\mathcal{E}\{\epsilon_i X_j\} = 0, \qquad i \neq j. \tag{1.4}$$

A second assumption, which is relaxed in models with correlated errors, is that the expected value of the product of the unique parts of two distinct items is zero,

$$\mathcal{E}\{\epsilon_i \epsilon_j\} = 0, \qquad i \neq j. \tag{1.5}$$

Given the first assumption, it can be shown (e.g., Bollen, 1989, p. 35) that Equation 1.3 implies that the population covariance matrix among the items $\boldsymbol{\Sigma}$ $(I \times I)$ can be written as a function of the loading matrix $\boldsymbol{\Lambda}$, LV covariance matrix $\boldsymbol{\Phi}$, and the error covariance matrix $\boldsymbol{\Psi}$,

$$\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}. \tag{1.6}$$

[1] The notation used throughout this dissertation is listed in Appendix B.

### 1.2.2 Models of Interest

In linear FA, or FA of the sample covariance matrix (FA-lin), item scores are assumed to be linearly related to the LV. Scales in the social sciences are often composed of items with a limited number of response categories. For example, the extent to which a respondent agrees with a statement is elicited, providing the choice between *not at all*, *not really*, *neutral*, *somewhat*, and *completely*. Such response categories are usually numerically coded as $\{0, 1, 2, 3, 4\}$. In practice, linear FA is often applied to investigate the psychometric properties of a set of such items. However, the assumption of a linear relationship between the item responses and the LV is violated by default then, because of the categorical nature of the items. Alternative approaches have been proposed, one of which is the use of estimated polychoric correlations instead of product-moment correlations.

Polychoric correlations (Olsson, 1979; Pearson & Pearson, 1922) can be estimated under the assumption of a normal continuous variable underlying the ordinal observed item response variable. The polychoric correlation reflects the association between these underlying variables. For each item, thresholds are estimated that link the underlying normal variable $X^*$ to the observed categorical variable $X$. To give an example, an $X^*$ value between threshold values $\tau_1$ and $\tau_2$ results in an $X$ value equal to 1, as illustrated in Figure 1.1.

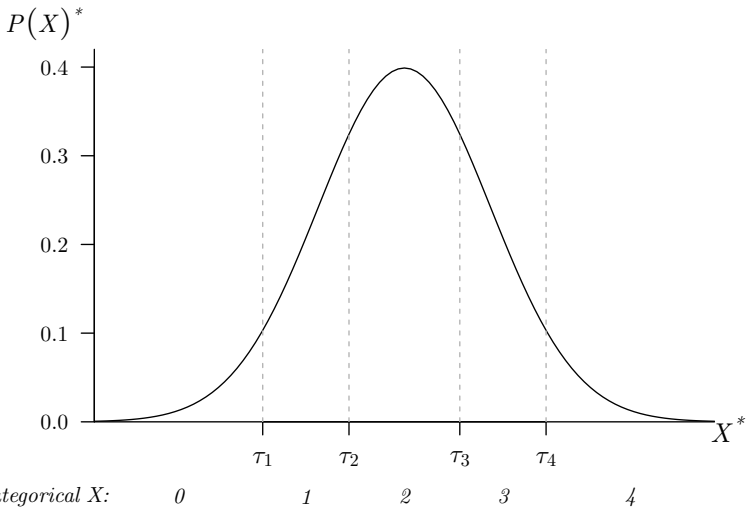



*Figure 1.1.* Illustration of a normal latent response variable $X^*$ underlying the five-point ordered categorical observed response variable $X$.

We refer to FA using polychoric correlations as FA-poly, which is defined as follows:

$$\mathbf{X}^* = \boldsymbol{\theta}\boldsymbol{\lambda}' + \mathbf{E}, \tag{1.7}$$
$$X_{is} = c - 1 \quad \text{for} \quad \tau_{i(c-1)} < X^*_{is} < \tau_{ic},$$

with $c = 1, 2, \ldots, C$, $\tau_{i0} = -\infty$, $\tau_{iC} = \infty$, $i = 1, 2, \ldots, I$, and $s = 1, 2, \ldots, n$. $\mathbf{X}^*$ $(n \times I)$ is a matrix of *latent* continuous item scores for respondents $s$ on items $i$, $\boldsymbol{\lambda}$ is a vector of item loadings of length $I$ on the latent dimension, $\boldsymbol{\theta}$ is the vector of LV scores of length $n$, $\mathbf{E}$ $(n \times I)$ is an error matrix with independent normally distributed elements $\epsilon_{is}$, $X_{is}$ are the *observed* categorical item scores, $\tau_{ic}$ are the thresholds of item $i$ for passing from category $c-1$ to $c$, and $C$ is the number of response categories.

In the remainder of our investigations, the focus is on FA-lin and FA-poly, where the former is included as a standard practice, and the latter as a theoretically more suitable alternative for the analysis of ordered categorical items.

### 1.2.3 Estimation

In this section, two estimation methods commonly employed in factor modeling are described briefly. In our studies, we employ maximum likelihood (ML) estimation for the linear FA model and a robust variant of weighted least squares (WLS) estimation for the polychoric FA model. Finally, the estimation of LV scores in factor modeling is discussed.

#### Maximum Likelihood Estimation

In FA model estimation, a fitting function is minimized with respect to the unknown parameters $\boldsymbol{\Lambda}$, $\boldsymbol{\Phi}$, and $\boldsymbol{\Psi}$. The most commonly used fitting function in linear FA estimation is ML. The ML fitting function $F_{\text{ML}}$ is defined as (e.g., Bollen, 1989, p. 107)

$$F_{\text{ML}} = \ln|\boldsymbol{\Sigma}(\boldsymbol{\xi})| + tr(\boldsymbol{\Sigma}(\boldsymbol{\xi})^{-1}\mathbf{S}) - \ln|\mathbf{S}| - I, \tag{1.8}$$

where $\boldsymbol{\Sigma}(\boldsymbol{\xi})$ is the model-implied covariance matrix given the model parameters $\boldsymbol{\xi}$, $|\cdot|$ is the matrix determinant, $tr(\cdot)$ refers to the trace of a matrix, $\mathbf{S}$ is the sample covariance matrix, and $I$ is the number of items.

Maximum likelihood estimators are asymptotically (for $n \to \infty$) (a) unbiased, (b) consistent, and (c) efficient. This means that, asymptotically, (a) the expectation of the parameter estimates equals the population value; (b) the probability of the estimator being arbitrarily close to the population value converges to one; and (c) the estimator has the smallest asymptotic variance among consistent estimators. Because these are asymptotic properties, they only hold approximately in large samples.

Equation 1.8 is derived under the assumption that the observed variables $X_i$ have a multivariate normal distribution or that $\mathbf{S}$ has a Wishart distribution (e.g., Bollen, 1989, p. 107), and only then is the estimator characterized by the three aforementioned properties. However, even under less restrictive assumptions the estimator has desirable properties. Asymptotic consistency, for example, does not

require the $X_i$'s to be multivariate normal; see Bollen (1989, p. 126ff.) for more detailed information.

### Weighted Least Squares Estimation

Weighted least squares estimation is usually done in three steps (B. O. Muthén, 1998–2004, p. 17ff.). In the first step, polychoric correlations are estimated in a two-stage procedure, where the thresholds $\tau_{ic}$ are estimated univariately before estimating the polychoric correlations given these thresholds (Olsson, 1979). Second, a weight matrix $\mathbf{W}$ is determined. In the final step, the model parameters are estimated by minimizing the fitting function

$$F_{\text{WLS}} = (\mathbf{r} - \boldsymbol{\rho}(\boldsymbol{\xi}))' \, \mathbf{W} \, (\mathbf{r} - \boldsymbol{\rho}(\boldsymbol{\xi})), \tag{1.9}$$

where $\mathbf{r}$ is a vector of estimated polychoric correlations, $\boldsymbol{\rho}(\boldsymbol{\xi})$ is a vector of model-implied correlations given the model parameters $\boldsymbol{\xi}$, and $\mathbf{W}$ is the weight matrix.

The weight matrix can be defined in various ways, leading to different estimation methods. In unweighted least squares (ULS) estimation, the weight matrix simply equals the identity matrix: $\mathbf{W} = \mathbf{I}$. In WLS estimation, $\mathbf{W}$ is the inverse of the estimated asymptotic covariance matrix of the polychoric correlation vector $\mathbf{r}$. In a robust type of WLS estimation (the default WLS in MPLUS when ordered categorical items are included in the analysis), only the diagonal elements of the WLS weight matrix are used. The mean adjusted WLS and mean-and-variance adjusted WLS (WLSMV) estimators from the MPLUS computer program (L. K. Muthén & Muthén, 1998–2010) are both robust WLS procedures and differ only in their definition of the $\chi^2$ goodness-of-fit statistic (B. O. Muthén, 1998–2004, p. 19). The diagonally WLS (DWLS) estimator in the LISREL computer program (Jöreskog & Sörbom, 1996) differs from WLSMV only in the computation of the estimated asymptotic covariance matrix $\mathbf{W}^{-1}$. Asymptotically, however, as $n \to \infty$ the two procedures are equivalent (e.g., Forero & Maydeu-Olivares, 2009).

### LV Score Estimation

In addition to the model parameters estimated by employing either ML or WLS estimation, it is possible to estimate LV scores $\theta_s$ for the sample of respondents using a separate procedure. In the MPLUS computer program, LV score estimation for FA-lin and FA-poly with WLS is performed using maximum a posteriori estimation, which makes use of the following form of Bayes' theorem (B. O. Muthén, 1998–2004, p. 47):

$$g(\theta_s|\mathbf{X}_s, \boldsymbol{\xi}) \propto f(\mathbf{X}_s|\theta_s) \, g(\theta_s), \tag{1.10}$$

i.e., the posterior distribution of LV score $\theta_s$ for respondent $s$, conditional on reponse pattern $\mathbf{X}_s$ and the item parameters $\boldsymbol{\xi}$, is proportional to the product of some function $f$ and the prior LV distribution $g(\theta_s)$, the latter being assumed to be normal.

For FA-lin, assuming continuous $X$ variables, function $f$ can be written as the product of function $f_i$ over all items $i$ (B. O. Muthén, 1998–2004, p. 48)

$$f_i(X_{is}|\theta_s) = \exp\left[-\frac{1}{2}(X_{is} - \lambda_i\theta_s)^2\psi_i^{-1}\right], \tag{1.11}$$

$$f(X_s|\theta_s) = \prod_{i=1}^{I} f_i(X_{is}|\theta_s), \tag{1.12}$$

where $X_{is}$ is the response of respondent $s$ to item $i$, $\lambda_i$ is the loading parameter for item $i$, and $\psi_i$ is the error variance of item $i$. Maximizing the log of the posterior with respect to $\theta_s$ then gives the same LV score estimates as obtained via the regression method (e.g., Bollen, 1989, p. 305)

$$\hat{\theta}_s = \hat{\boldsymbol{\lambda}}'\,\hat{\boldsymbol{\Sigma}}^{-1}\,\mathbf{X}_s, \tag{1.13}$$

where $\hat{\boldsymbol{\lambda}}$ $(I \times 1)$ is the vector of estimated loading parameters and $\hat{\boldsymbol{\Sigma}}$ $(I \times I)$ is the estimated covariance matrix among the items.

For categorical observed variables, as in FA-poly, $f$ is obtained as (B. O. Muthén, 1998–2004, p. 48)

$$f_i(X_{is}|\theta_s) = \Phi[(\tau_{i(c+1)} - \lambda_i'\theta_s)\,\psi_i^{-1/2}] - \Phi[(\tau_{ic} - \lambda_i'\theta_s)\,\psi_i^{-1/2}], \tag{1.14}$$

where $\Phi$ is the standard normal cumulative distribution function and $\tau_{ic}$ is the threshold parameter for category $c$ of item $i$. The LV score estimate is found as the maximum of the posterior $\theta$ distribution, by minimizing the function

$$F = \frac{1}{2}\,(\theta_s - \mu)'\,\sigma^{-1}(\theta_s - \mu) - \sum_{i=1}^{I} \ln\, f_i(X_{is}|\theta_s) \tag{1.15}$$

with respect to $\theta_s$, where $\mu$ and $\sigma^2$ are the mean and the variance, respectively, of the normal prior distribution, respectively, and ln denotes the natural logarithm. In MPLUS quasi-Newton techniques are used to iteratively perform the minimization of $F$, using only first-order derivatives of $F$ to the unknown parameters (B. O. Muthén, 1998–2004, p. 48).

## 1.3 Item Response Theory

Independent of the FA developments, another line of scale construction was initiated by Guttman (1945). The focus of IRT modeling is on the relation between (individual) item responses and (individual) latent trait values, represented by an item response function (IRF). Guttman scaling is the deterministic precursor of nonparametric IRT (NIRT; Mokken, 1971; I. W. Molenaar, 1991). The key feature of NIRT is the nonparametric definition of (nondecreasing) IRFs and the concept of scalability based on homogeneity.

Two parametric developments of IRT were initiated in the 1950s and 1960s, complementary to the nonparametric models. Focusing on the scientific properties of measurement models, Rasch (1960) developed a family of IRT models, which were later extended by scholars like David Andrich and Geoffrey Masters. The other line of parametric IRT development started with Lord and Novick's (1968) classical textbook, in which the two-parameter IRT model was introduced.

Some distinctions should be addressed with regard to IRT models. The first, already mentioned, distinction is between parametric and nonparametric IRT models. In nonparametric IRT models, the shape of the IRFs is less restricted than in parametric IRT models. Parametric models are far more popular in practice than nonparametric models, and, consequently, a wide variety of models is in that category.

Additionally, one can distinguish between dichotomous (two response categories) and poly(cho)tomous (more than two response categories) models. Most IRT models were originally developed for dichotomous items, and later extended to polytomous items.

A final distinction is that between unidimensional and multidimensional models, taking one or multiple LVs into account, respectively. Some unidimensional IRT models have been extended to include multiple dimensions. Multidimensional IRT models are not discussed here, as they are beyond the scope of this dissertation; the interested reader is referred to Reckase (2009).

### 1.3.1 Definitions and Assumptions

In IRT, the *probability* that a respondent endorses an item is modeled, where endorsement of an item means scoring positively, e.g., scoring 1 on a dichotomous $\{0, 1\}$ item. For each item, that probability is modeled as a function of the LV, resulting in a monotone nondecreasing IRF, as it is assumed that the probability increases with the LV value.[2]

Except for the Guttman model, IRT models are not deterministic. In general, it is assumed that respondent $s$, who is more able than respondent $t$, is *more likely* to endorse an item than respondent $t$. It is also assumed that it is more likely that a respondent endorses an easier item than a more difficult item. The probabilistic IRT model allows for unlikely response patterns of less able respondents scoring higher than more able respondents due to sources of error left unspecified.

Since the probability of endorsing an item is modeled, the range of the function that describes the probability should be restricted to the $[0, 1]$ range. The cumulative normal distribution function and the logistic function are common choices for modeling the probability.

[2] In most models, endorsing an item is a sign of a higher score on the LV, and the more items are endorsed, the higher the LV score. This assumption is central to cumulative models. The class of IRT models known as unfolding or proximity models, in which the probability of endorsing an item first increases with the LV, but at a certain point decreases (e.g., Andrich, 1997) is beyond the scope of this dissertation.

One of the best-known IRT models is IRT-2p, where the IRF is modeled as

$$P(X_{is} = 1|\theta_s, \alpha_i, \beta_i) = \frac{\exp[\alpha_i(\theta_s - \beta_i)]}{1 + \exp[\alpha_i(\theta_s - \beta_i)]}, \tag{1.16}$$

which says that the probability that an item $i$ is endorsed by respondent $s$ depends on the respondent's LV score $\theta_s$, the item discrimination $\alpha_i$, and the item difficulty $\beta_i$. The difficulty parameter determines the location of the IRF on the LV scale and is defined as the LV value for which the probability of endorsement is equal to 0.5. Respondents with LV scores higher than the difficulty of an item are more likely to endorse the item than not to endorse it.

The discrimination parameter indicates the degree to which an item can distinguish between respondents with different LV scores. A large discrimination parameter results in a steep IRF. Item discrimination is a local property: Highly discriminative items can discriminate well between respondents with LV scores around the item difficulty, but poorly between respondents who are located further along either side of the scale. In the Rasch model, all discrimination parameters are restricted to be equal. For illustrative purposes, three IRFs with discrimination and difficulty parameter values are shown in Figure 1.2.

There are two distinct assumptions in IRT: local independence and the specific form of the IRFs (Embretson & Reise, 2000, p. 45ff.). Local independence means that each common factor underlying the set of items is specified as an LV, and conditionally on the LV(s), there is no covariance between the items, or, more formally,

$$P(X_{1s}, X_{2s}, \ldots, X_{Is}|\boldsymbol{\theta_s}) = \prod_{i=1}^{I} P(X_{is}|\boldsymbol{\theta_s}), \tag{1.17}$$

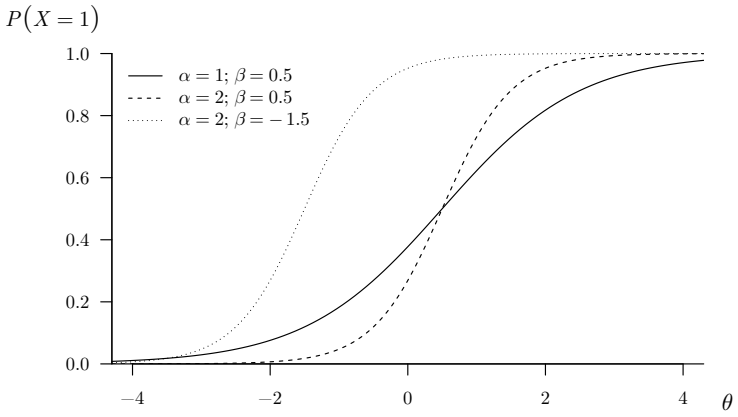



*Figure 1.2.* Three item response functions with various discrimination ($\alpha$) and difficulty ($\beta$) parameters.

where $\boldsymbol{\theta_s}$ is an element of the specified latent space, i.e., all $Q$ latent traits $(\theta_{s1}, \theta_{s2}, \ldots, \theta_{sQ})$. Lord and Novick (1968, p. 361) show that the assumption of local independence is equivalent to a complete specification of the latent space. They show that when Equation 1.17 is divided by

$$P(X_{2s}, X_{3s}, \ldots, X_{Is}|\boldsymbol{\theta_s}) = \prod_{i=2}^{I} P(X_{is}|\boldsymbol{\theta_s}), \tag{1.18}$$

it holds that

$$P(X_{1s}|\boldsymbol{\theta_s}; X_{2s}, X_{3s}, \ldots, X_{Is}) = P(X_{1s}|\boldsymbol{\theta_s}). \tag{1.19}$$

Naturally, this generalizes to all other items. If Equation 1.19 does not hold, the conditional probability of endorsing $X_1$ for respondents with the same $\boldsymbol{\theta}$ will be different, depending on their scores on $X_2, X_3, \ldots, X_I$. This cannot occur when the complete latent space is specified. So, local independence implies a specification of the complete latent space, and vice versa. When the latent space is made up of only one latent trait, local independence implies unidimensionality, and vice versa. However, in the literature (e.g., Bejar, 1983; Finger, 2001; Hambleton & Swaminathan, 1985; Hulin, Drasgow, & Parsons, 1983; Sijtsma & Molenaar, 2002) the concepts of local independence and unidimensionality are often introduced as separate assumptions without mentioning their interdependence.

The second assumption in IRT concerns the IRFs. An IRF describes the relationship between the response to an item and the LV. Different IRT models specify different mathematical forms for the IRFs. It is assumed that the shape of the IRFs is correctly specified (Embretson & Reise, 2000, p. 45), so the applicable assumption depends on the specification of the IRF in each model.

### 1.3.2 Models of Interest

There are many IRT models, differing in the way the IRFs are defined mathematically. For a comprehensive overview, the reader is referred to Van der Linden and Hambleton (1997). We refer to Van der Ark (2001) for a clarification of the relationships among a large number of IRT models. In this section, we present the IRT models focused on in the remainder of our investigations.

Our research is aimed at ordered categorical items, thus polytomous models are required. A widely applied model is Samejima's (1969) polytomous extension of the IRT-2p: the graded response IRT model (IRT-grm). The *normal-ogive* version of the model is defined as

$$P(X_{is} = c \,|\, \theta_s, \alpha_i^{\mathcal{N}}, \beta_{ic}) = \frac{1}{\sqrt{2\pi}} \int_{\alpha_i^{\mathcal{N}}(\theta_s - \beta_{i(c+1)})}^{\alpha_i^{\mathcal{N}}(\theta_s - \beta_{ic})} e^{-t^2/2} dt, \tag{1.20}$$

where $P(X_{is} = c|\ldots)$ denotes the conditional probability of respondent $s$ choosing category $c$ of item $i$, $\theta_s$ is the respondent's LV score, $\alpha_i^{\mathcal{N}}$ is the discrimination parameter in the normal-ogive model, and $\beta_{ic}$ is the step-difficulty parameter.

The *logistic* version of the model (also presented in Samejima, 1969) is more commonly applied in the behavioral and social sciences:

$$P(X_{is} \geq c \,|\, \theta_s, \alpha_i, \beta_{ic}) = \frac{1}{1 + \exp[-\alpha_i(\theta_s - \beta_{ic})]}, \tag{1.21}$$

where $\alpha_i$ is the discrimination parameter for item $i$ in the logistic model. The IRT-grm discrimination or slope parameter does not differ from the IRT-2p discrimination parameter. The IRT-2p difficulty parameter, however, is replaced by a step-difficulty parameter. Instead of one response function for each item, IRT-grm defines one response function per item step. So for a five-category item, four item-step response functions (ISRFs) are in place. An item is thus treated as a series of dichotomies, all with estimated IRT-2p models constrained to have equal discrimination parameters (Embretson & Reise, 2000, p. 99).

The second IRT model we focus on is a nonparametric model. In the monotone homogeneity model, introduced by Mokken (1971), the only restriction on the IRFs is that they are monotonically nondecreasing. It is included in our comparative studies because, due to its weak assumptions, this model is widely applicable and still produces quite useful results in terms of scale construction and evaluation. Furthermore, comparisons between NIRT and FA are scarce.

The Mokken model was extended to polytomous items by I. W. Molenaar (1991) in a way similar to the extension of IRT-2p to IRT-grm. In the nonparametric Mokken IRT model (IRT-mok) a polytomous item is also taken to be a series of $(C-1)$ dichotomies, with $C$ the number of response categories.

Two distinct assumptions are posed in IRT-mok: monotonicity and local independence. Monotonicity refers to the restriction as was mentioned before on the LVs of being monotonically nondecreasing. The local independence assumption is comparable to the parametric variant, i.e., conditional on the LV, item response variables are independent. As the IRT-mok model accounts for only one LV, the local independence assumption is equal to the assumption of unidimensionality here, i.e., the complete latent space is covered by a single LV.

### 1.3.3 Estimation

In this section the estimation methods employed in our research are described. For the IRT-grm model, marginal ML (MML) as applied to the IRT-2p model is described, as it is more comprehensible in that form and generalizes to the polytomous case. The likelihood equations for the model parameters are given in quite some detail, following Baker and Kim (2004), but rewritten in the notation employed in this dissertation. In this way, we aim to provide a comprehensive overview of the estimation methods applied, facilitating the comparison with the FA estimation methods.

For IRT-mok, the term model estimation is perhaps questionable, as the Mokken model does not aim to estimate any population parameters. However, one can define a population parameter of the scaling coefficient known as Loevinger's $H$, based on

FA population parameters as we will show in Chapter 4, and one can estimate it in a sample.

### Marginal Maximum Likelihood

Whereas in FA modeling factor scores are usually of secondary interest and receive relatively little attention, in the IRT tradition person parameters are equally important as item parameters and LV scores are included in the model estimation process by default.

For MML to be employed for IRT model estimation, it must be assumed that the respondents are a random sample from a population in which the LV is distributed according to a specific density function (Baker & Kim, 2004, p. 158). Usually, the normal density function is used, but in principle, any distribution can be employed; the density function can even be estimated from the data to incorporate deviating distributional shapes, such as severe skewness.

To separate the estimation of item parameters from the estimation of person parameters, one integrates over the LV distribution, thus removing the person parameters from the likelihood equation. The item parameters are estimated in the marginal distribution. Since the estimation of item parameters only depends on the LV distribution, the number of parameters does not increase with the sample size, leading to consistent item parameter estimators.

For the IRT-2p, Baker and Kim (2004, p. 160ff) show that the Bock and Lieberman (1970) solution for the marginal likelihood equation of the discrimination parameter $\alpha$ is

$$\frac{\partial \ln L}{\partial \alpha_i} = \sum_{s=1}^{n} \int [X_{is} - P_i(\theta_s)](\theta_s - \beta_i)[P(\theta_s|\mathbf{X}_s, \boldsymbol{\xi}, \boldsymbol{\eta})]\, d\theta_s, \tag{1.22}$$

where

$$P(\theta_s|\mathbf{X}_s, \boldsymbol{\eta}, \boldsymbol{\xi}) = \frac{P(\mathbf{u}_s|\theta_s, \boldsymbol{\xi})\, g(\theta_s|\eta)}{\int P(\mathbf{u}_s|\theta_s, \boldsymbol{\xi})\, g(\theta_s|\eta)\, d\theta_s}, \tag{1.23}$$

in which

$$P(\mathbf{u}_s|\theta_s, \boldsymbol{\xi}) = \prod_{i=1}^{I} P_i(\theta_s)^{X_{is}}\, [1 - P_i(\theta_s)]^{1-X_{is}}, \tag{1.24}$$

and where $X_{is}$ is the reponse of respondent $s$ to item $i$, $P_i(\theta_s)$ is the probability of endorsing item $i$ given LV score $\theta_s$, $\mathbf{X}_s$ is the vector of item responses of respondent $s$, $\boldsymbol{\xi}$ indicate the model parameters $\alpha$ and $\beta$, and $\boldsymbol{\eta}$ is the vector of parameters of the LV distribution. Baker and Kim go on to prove the likelihood equation for the difficulty parameter $\beta_i$ to be

$$\frac{\partial \ln L}{\partial \beta_i} = -\alpha_i \sum_{s=1}^{n} \int [X_{is} - P_i(\theta_s)][P(\theta_s|\mathbf{X}_s, \alpha_i, \beta_i, \boldsymbol{\xi})] d\theta_s = 0. \tag{1.25}$$

The reformulation of these likelihoods by Bock and Aitken (1981) employing approximations to the integrals using Hermite-Gauss quadrature give the marginal likelihood equations as follows (Baker & Kim, 2004, p. 166ff.)

$$\alpha_i : \quad \sum_{v=1}^{V}(T_v - \beta_i)[h_{iv} - f_{iv}\, P(T_v)] = 0, \tag{1.26}$$

$$\beta_i : \quad -\alpha_i \sum_{v=1}^{V}[h_{iv} - f_{iv}\, P(T_v)] = 0, \tag{1.27}$$

with

$$f_{iv} = \sum_{s=1}^{n}\left[\frac{\prod_{i=1}^{I} P_i(T_v)^{X_{is}}[1 - P_i(T_v)]^{1-X_{is}}\, w_{T_v}}{\sum_{v=1}^{V}\prod_{i=1}^{I} P_i(T_v)^{X_{is}}[1 - P_i(T_v)]^{1-X_{is}}\, w_{T_v}}\right], \tag{1.28}$$

and

$$h_{iv} = \sum_{s=1}^{n}\left[\frac{\prod_{i=1}^{I} X_{is}P_i(T_v)^{X_{is}}[1 - P_i(T_v)]^{1-X_{is}}\, w_{T_v}}{\sum_{v=1}^{V}\prod_{i=1}^{I} P_i(T_v)^{X_{is}}[1 - P_i(T_v)]^{1-X_{is}}\, w_{T_v}}\right], \tag{1.29}$$

since

$$P(T_v|\mathbf{X}_s, \boldsymbol{\xi}, \boldsymbol{\eta}) = \frac{\prod_{i=1}^{I} P_i(T_v)^{X_{is}}[1 - P_i(T_v)]^{1-X_{is}}\, w_{T_v}}{\sum_{v=1}^{V}\prod_{i=1}^{I} X_{is}P_i(T_v)^{X_{is}}[1 - P_i(T_v)]^{1-X_{is}}\, w_{T_v}}, \tag{1.30}$$

and

$$P_i(T_v) = \frac{\exp[\alpha_i(T_v - \beta_i)]}{1 + \exp[\alpha_i(T_v - \beta_i)]}. \tag{1.31}$$

Here $T_v$, $v = 1, 2, \ldots, V$, is a node — or midpoint of a rectangle — on the LV scale divided into $V$ discrete parts, $w_{T_v}$ is the weight associated with $T_v$ based on the density function $g(\theta|\eta)$. Furthermore, $f_{iv}$ is the expected number of respondents of LV level $T_v$ attempting item $i$, and $h_{iv}$ is the expected number of endorsements for item $i$ at node $v$.

Because $f_{iv}$ and $h_{iv}$ depend on the values of the item parameters, they cannot be estimated simultaneously with the item parameters. As a solution, the iterative expectation maximization (EM) algorithm (Dempster, Laird, & Rubin, 1977) is employed. In IRT-2P the expectation and maximization steps are implemented as follows (cf. Baker & Kim, 2004, p. 171). In the expectation step, first the likelihood of each response vector is computed at each of the $V$ nodes using provisional item parameters and

$$L(T_v) = \prod_{i=1}^{I} P_i(T_v)^{X_{is}}\,[1 - P_i(T_v)]^{1-X_{is}}, \tag{1.32}$$

representing the quadrature form of $P(\mathbf{X}_s|\theta_s = T_v, \boldsymbol{\xi})$. Next, the posterior probability that the LV score of respondent $s$ equals $T_v$ is calculated using the quadrature weights

$w_{T_v}$ and

$$P(T_v|\mathbf{X}_s, \boldsymbol{\xi}, \boldsymbol{\eta}) = \frac{L(T_v)\, w_{T_v}}{\sum_{v=1}^{V} L(T_v)\, w_{T_v}}. \tag{1.33}$$

And finally, $f_{iv}$ and $h_{iv}$ are computed.

In the maximization step, likelihood Equations 1.26 and 1.27 are solved for the item parameters $\alpha_i$ and $\beta_i$ using the estimates of $f_{iv}$ and $h_{iv}$. As $f_{iv}$ and $h_{iv}$ depend on the item parameters, the likelihood equations have to be solved iteratively.

In an additional step, the likelihood of each parameter is compared to the likelihood from the previous cycle. If there is no change — or if the difference is smaller than a certain criterion — the process has converged and is terminated. Otherwise, a new EM cycle is initiated.

It should be noted that MML item parameter estimation is not possible for items that are endorsed by none or all of the respondents in the sample, because they would be infinite.

### LV Score Estimation

LV scores in IRT modeling can be estimated using a number of methods, one of which is expected a posteriori (EAP) estimation. EAP is based on Bayes' theorem in the following form (cf. Baker & Kim, 2004, p. 193)

$$g(\theta_s|\mathbf{X}_s, \boldsymbol{\xi}) = \frac{P(\mathbf{X}_s|\boldsymbol{\theta}_s, \boldsymbol{\xi})\, g(\boldsymbol{\theta})}{\mathbf{X}_s}. \tag{1.34}$$

In quadrature form, the expected LV score for respondent $s$, given response pattern $\mathbf{X}_s$ and LV distribution parameters $\boldsymbol{\eta}$, can be determined noniteratively using

$$\mathcal{E}(\theta_s|\mathbf{X}_s, \boldsymbol{\xi}) = \frac{\sum_{v=1}^{V} T_v\, L(T_v)\, w_{T_v}}{\sum_{v=1}^{V} L(T_v)\, w_{T_v}}, \tag{1.35}$$

where both the quadrature weights $w_{T_v}$ and likelihoods $L(X_s|T_v)$ can be taken from the final stage of the item parameter estimation procedure.

### Robust Maximum Likelihood

Using the MPLUS computer program, the IRT-grm model can be estimated by a robust ML (MLR) procedure. This is a generalization of the MML-EM procedure for IRT-2p to polytomous items. The difference with standard ML concerns only the estimation of standard errors and $\chi^2$ values. Standard errors are computed using a sandwich estimator and the $\chi^2$ estimates are asymptotically equivalent to Yuan and Bentler's (2000) $T_2^*$ statistic (L. K. Muthén & Muthén, 1998–2010, p. 484). LV scores are estimated using the aforementioned EAP method.

**Loevinger's $H$ coefficient**

A useful result of IRT-mok is the scalability coefficient known as Loevinger's $H$, which can be computed and well interpreted if the assumptions of monotonicity and local independence are satisfied. Loevinger's $H_{ij}$ for item pair $(i, j)$ is directly related to the concept of a Guttman error, i.e., passing an item step that requires a certain LV score, while failing an item step that demands a lower LV score of the respondent, and is defined as (I. W. Molenaar, 1991)

$$H_{ij} = 1 - \frac{F_{ij}}{E_{ij}}, \tag{1.36}$$

where $F_{ij}$ is the total weighted number of Guttman errors for item pair $(i, j)$ and $E_{ij}$ is the expected total weighted number of Guttman errors if the items were independent. The weighting is done by the number of item steps corresponding to the two items for which the ordering is erroneous (see I. W. Molenaar, 1991, for more information). Hence, both $F_{ij}$ and $E_{ij}$ can be computed based on the bivariate frequency table of items $i$ and $j$.

From the pairwise $F_{ij}$ and $E_{ij}$ values, $H_i$ and $H_{scale}$ can be computed as follows:

$$H_i = 1 - \frac{\sum_{\substack{j=1 \\ j \neq i}}^{I} F_{ij}}{\sum_{\substack{j=1 \\ j \neq i}}^{I} E_{ij}}, \tag{1.37}$$

$$H_{scale} = 1 - \frac{\sum_{i}^{I} \sum_{\substack{j=1 \\ j \neq i}}^{I} F_{ij}}{\sum_{i}^{I} \sum_{\substack{j=1 \\ j \neq i}}^{I} E_{ij}}. \tag{1.38}$$

The scalability coefficient provides a means of quantifying the strength of an item pair ($H_{ij}$), an item ($H_i$), or a scale ($H_{scale}$). As a rule of thumb, a set of items is only considered to be a scale when $H_{scale} \geq 0.3$, and that scale is considered weak when $0.3 \leq H_{scale} < 0.4$, medium when $0.4 \leq H_{scale} < 0.5$, and strong when $H_{scale} \geq 0.5$ (Sijtsma & Molenaar, 2002, p. 60).

## 1.4 Comparison of FA and IRT

In this section, a theoretical comparison is made between FA and IRT. We compare the model formulas and demonstrate the relations between FA and IRT parameters, following Takane and De Leeuw (1987). Subsequently, the main distinctions in estimation procedures of the two groups of models are brought to the reader's attention.

### 1.4.1 Comparison of FA and IRT Parameters

Takane and De Leeuw (1987) formally proved the equivalence of the *two-parameter normal-ogive model* and the *polychoric common factor model* for both dichotomous

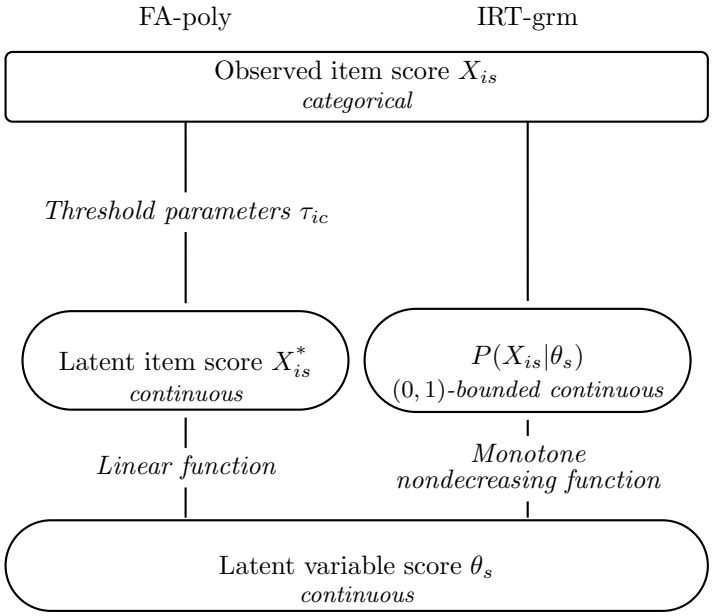



*Figure 1.3.* Graphical representation of FA-poly and IRT-grm.

and polytomous items. The equivalence approximately holds for the two-parameter logistic model, to the extent that the logistic distribution approximates the normal distribution. Takane and De Leeuw stated that IRT and FA are two alternative formulations of the same model. The difference lies at which point in the estimation procedure the marginalization is performed. Moreover, the commonly employed estimation methods differ between FA and IRT, adding to the difference in results.

In FA the marginalization is performed on the latent continuous item variables $X_i^*$ first, and the categorization is performed next. In practice, polychoric correlations between the latent continuous item variables are computed first. This requires estimation of the thresholds $\tau_{ic}$. The polychoric correlation matrix is sufficient for subsequent computations.

In IRT the order is reversed: First the categorization of the latent response variable is done conditionally on $\theta$, and next the marginalization. The categorization, however, is done implicitly, as the latent continuous item variable is not explicitly part of the IRT model. Instead, the observed scores are used directly to compute the conditional probability. Subsequently, a marginal probability can be computed, integrating over $\theta$.

The links between the observed item scores and the LV scores for FA-poly and IRT-grm are depicted in Figure 1.3. The starting point of an analysis is usually a data

matrix with observed item scores $X_{is}$. The objective is to obtain LV estimates $\theta_s$ for the respondents based on their response patterns. The left part of the figure illustrates that in FA polychoric correlations are estimated, which map the categorical observed item scores $X_{is}$ to continuous latent item scores $X_{is}^*$ by estimating thresholds $\tau_{ic}$. The $X_{is}^*$ variables are modeled to be related linearly to the LV. As a result, LV scores can be estimated.

On the right side of Figure 1.3 it is shown that in IRT the modeling starts by estimation of $P(X_{is}|\theta_s)$, which is a continuous function bounded between 0 and 1. This estimation is directly based on the categorical observed item scores. Subsequently, LV estimates are obtained, since $P(X_{is}|\theta_s)$ is a function of the LV.

When it is assumed that $\theta \sim \mathcal{N}(\mu, \sigma^2)$ and $\epsilon_i \sim \mathcal{N}(0, \psi_i)$, the conditional probability that $X_i^*$ is less than or equal to the threshold $\tau_{ic}$ for category $c$, given the LV score $\theta_s$, can be expressed as (cf. Mehta & Taylor, 2006)

$$P(X_i^* \leq \tau_{ic}|\theta_s) = \Phi\left[\frac{\tau_{ic} - \lambda_i\theta_s)}{\sqrt{\boldsymbol{\psi_i}}}\right] = \Phi\left[\frac{\lambda_i}{\sqrt{\psi_i}}\left(\frac{\tau_{ic}}{\lambda_i} - \theta_s\right)\right]. \tag{1.39}$$

The conditional probability for the normal-ogive IRT-grm model can be written as

$$P(X_i \leq c|\theta_s) = \Phi\Big[\alpha_i^{\mathcal{N}}(\beta_{ci} - \theta_s)\Big]. \tag{1.40}$$

When Equation 1.39 is compared with Equation 1.40, the following relations between the IRT and FA parameters become apparent:

$$\alpha_i^{\mathcal{N}} = \frac{\lambda_i}{\sqrt{\psi_i}}, \tag{1.41}$$

and

$$\beta_{ic} = \frac{\tau_{ic}}{\lambda_i}. \tag{1.42}$$

These equations hold for the *normal-ogive* IRT-grm. For the more commonly applied *logistic* version of this model, an adjustment to Equation 4.4 is required. The normal-ogive and logistic-ogive function are nearly equivalent, when a scaling factor $d \equiv 1.702$ is used: $\alpha_i = d\,\alpha_i^{\mathcal{N}}$ (see, e.g., Camilli, 1994; I. W. Molenaar, 1974). Thus, to accommodate the logit scale of the logistic IRT-grm, we use

$$\alpha_i = d\frac{\lambda_i}{\sqrt{\psi_i}}, \tag{1.43}$$

where $\alpha_i$ is the discrimination parameter for item $i$ in the logistic IRT-grm model.

For model identification, additional restrictions are required on the latent continuous item variables $X^*$ and/or the LV $\theta$, for which several possibities exist leading to different model parameterizations. Traditionally, FA and IRT each have their own distinct parameterizations. Model identification in FA is ensured by either fixing one of the loadings or fixing the LV variance, usually to 1. In IRT the error variance is

set to 1. In IRT software the latter is usually done implicitly by taking the standard normal distribution function to model the error variables.

In addition to the identification constraints, parameters are often further constrained to provide standardized parameters, which have the advantage of being independent of the metrics of the LV, the latent continuous item variables, and the observed items. FA parameters are often standardized to unit latent continuous item variance (in addition to the loading or LV variance constraint). IRT parameters are often standardized to unit LV variance (in addition to the error variance constraint).

These different parameterization and standardization defaults for FA and IRT complicate the comparison of the results of these respective models in practice slightly and call for careful consideration when comparing FA and IRT parameter estimation results.

### 1.4.2 Comparison of FA and IRT Estimation

In the literature comparing FA and IRT, often a distinction is made between limited-information (LI) and full-information (FI) estimation. FA methods are usually associated with LI, and IRT methods are often referred to as FI (e.g., Bolt, 2005).

In the edited volume *Contemporary Psychometrics* (Maydeu-Olivares & McArdle, 2005), an entire chapter was devoted to FI and LI estimation of IRT and FA models (Bolt, 2005). According to Bolt, LI and FI differ in the amount of information that is used from the data matrix of item responses. For example, when item responses on a test of $I$ five-category items are put in a contingency table, we obtain an $I$-dimensional table with $5^I$ cells containing each possible combination of item responses. FI estimation methods use all of these cell frequencies and thus try and fit a model considering all the observed response patterns. LI estimation methods take only marginal frequencies of the contingency table into account (collapsing over dimensions), usually up to the second-order margins, i.e., univariate and bivariate item information.

Christoffersson (1975) introduced a generalized least squares (GLS) approach to FA parameter estimation based on the univariate and bivariate marginal item distributions, and compared it to unconditional ML, i.e., ML based on all information in the data, noting that the former involves a *loss of information* compared to the second. He showed, however, that this loss of information, due to ignoring all joint probabilities higher than the first and second order, is of no practical consequence. Bock, Gibbons, and Muraki (1988) referred to Christoffersson's (1975) GLS method as LI, because it is based on "low-order joint occurrence frequencies of the item scores," in contrast to FI that uses "the frequencies of all distinct item response vectors" (Bock et al., 1988, p. 262). Various authors (e.g., Boulet, 1996; Jöreskog & Moustaki, 2001; Knol & Berger, 1991) consent to this terminology[3].

[3]One notable exception can be found in Verschuren (1991, p. 219ff.) who distinguished between LI and FI in terms of the *number of estimation steps*. According to his terminology, in LI parameters are estimated per equation, whereas in FI all parameters are estimated simultaneously. We consider this, however, as an exceptional use of terminology.

Directly related to the distinction between FI and LI are two definitions of local independence (see also McDonald, 1981, 1997). Since FI uses each individual response pattern, the *strong* principle of local independence holds, i.e.,

$$P(\mathbf{X}_s|\theta_s) = \prod_{i=1}^{I} P(X_{is}|\theta_s), \tag{1.44}$$

which means that the probability of observing response vector $\mathbf{X}_s$ for respondent $s$ with an LV score $\theta_s$ can be expressed as the product of the probabilities of observing each conditional item score $X_{is}$ in the response vector. LI methods, on the other hand, are connected to the *weak* principle of local independence, which holds when the covariances of all item pairs $(i, j)$ conditional on the LV are equal to zero,

$$cov(X_i, X_j|\boldsymbol{\theta}) = 0. \tag{1.45}$$

LI estimation methods only take the covariances of the items into account, which implies the use of the first- and second-order marginals.

In conclusion, LI and FI are generally distinguished by the information taken into account, with LI using only the univariate and bivariate frequencies, and FI using the complete $I$-variate response patterns as a basis for parameter estimation. Thus, in general, IRT methods use more information from the raw data matrix $\mathbf{X}$ than FA methods do.

## 1.5 Conclusion

In this chapter the two approaches to scale analysis under investigation, factor analysis (FA) and item response theory (IRT), were introduced. From these two independently developed traditions, scaling models evolved that bear a great number of similarities. Mathematically, some FA and IRT models are even equivalent.

FA and IRT also have their respective traditions in estimation methods applied, with FA linked to limited-information (LI) and IRT connected to full-information (FI) methods.

In the chapters to follow we focus on the following models: FA of the sample covariance matrix by means of maximum likelihood (FA-lin-ML), FA of the estimated polychoric correlation matrix by means of mean-and-variance adjusted weighted least squares (FA-poly-WLSMV), the graded response model by means of robust maximum likelihood (IRT-grm-MLR), and the nonparametric Mokken IRT model (IRT-mok). FA of the sample covariance matrix (FA-lin) is included as the standard practice, as will be corroborated in the next chapter. It should be noted that, in spite of availability of robust ML estimators available, we chose to include the standard ML being the most commonly applied practice of FA modeling. FA of the estimated polychoric correlation matrix (FA-poly) and the graded response IRT model (IRT-grm) are of interest, because they are theoretically the most appropriate for the analysis of scales consisting of ordered categorical (i.e., Likert) items, as commonly used in the social and behavioral

sciences. Additionally, these three parametric models are compared to IRT-mok, a nonparametric model, posing fewer restrictions on the data.

We will further investigate whether in the practice of scale analysis — characterized by finite samples and nonnormally distributed data — the FI methods typically applied in IRT-grm hold an advantage over FA-poly, traditionally connected to LI estimation. But before doing so, we first turn to the *practice* of scale analysis. In the next chapter, we therefore review the application of scaling models as practiced by applied researchers, primarily interested in characteristics of people as measured by means of questionnaires or scales.

# Chapter 2

# Comparison of FA and IRT in Practice

In the process of scale construction and evaluation, statistical modeling is used to assess the extent to which a group of items can be considered to measure the latent variable(s) researchers are interested in. Factor analysis (FA) and item response theory (IRT) are two types of models used for scale analysis. In the study presented in this chapter, we aim to evaluate the use of FA and IRT for the construction and evaluation of scales *in practice*.

Of primary interest is the researchers' motivation for choosing either methodology. Of secondary (methodological) interest is to investigate how the analysis is performed and which of its results are reported.

As was pointed out in the previous chapter, the theoretical relationship between FA and IRT has been well documented (Takane & De Leeuw, 1987; see also Kamata & Bauer, 2008, and Mehta & Taylor, 2006), demonstrating that certain variants of FA and IRT are equivalent, thus enabling the computation of FA model parameters from IRT parameters and vice versa (see T. A. Brown, 2006, p. 398ff., for a comprehensive demonstration). Furthermore, in previous research, the results of FA and IRT analysis have been compared by means of Monte Carlo studies. Before turning to an overview of that research in the next chapter, we first focus on what the *practice* of scale development is: What method is used, why is it used, and how is it used? We thus provide a descriptive overview of the current status of scale construction and evaluation. We shall also consider whether and where there is room for improvement.

We conducted a review of a sample of published articles. We first searched for journals in the fields of psychology and education that contain many articles reporting on the construction or evaluation of a scale as the main topic. *Psychological Assess-*

This chapter is a slightly adapted version of Ten Holt, J.C., Van Duijn, M.A.J. & Boomsma, A. (2010). Scale construction and evaluation in practice: A review of factor analysis versus item response theory applications. *Psychological Test and Assessment Modeling*, *52*, 272–297.

*ment* (PA), *European Journal of Psychological Assessment* (EJPA), and *Educational and Psychological Measurement* (EPM) met this criterion. In these journals, all articles concerning scale construction or evaluation published in 2005 were selected for review, which amounted to 46 articles. In the majority of articles either a factor or IRT analysis was conducted. Six articles that contained other analyses were excluded from this review, due to their lack of relevance for the present comparison. Three of these articles concerned a reliability generalization analysis (cf. Vacha-Haase, 1998). In the other three cases, a classical test theory (CTT) analysis, a generalizability analysis, and a multidimensional scaling analysis were performed, respectively. Of the remaining 40 articles, one contained separate analyses of two distinct scales, which were both included. Hence, a total of 41 studies were selected for investigation.

First, we describe how often each model is applied. Second, the extent to which explicit motivation was given for model choice is discussed. Third, an attempt is made to reveal implicit motives by examining various characteristics of the data and the models. Fourth, we review how the statistical analyses were performed and reported upon. Finally, the findings of our study are summarized and discussed.

## 2.1 Model Choice

FA, IRT, and both FA and IRT are applied in 32 (78%), six (15%), and three (7%) studies, respectively, illustrating the dominance of FA in the practice of scale analysis. These percentages and all those mentioned hereafter refer to all 41 studies included in the review.

In four FA studies, the applied model has an equivalent IRT counterpart. In these studies, the model is estimated using polychoric correlations, which is equivalent (cf. Takane & De Leeuw, 1987) to using the two-parameter IRT model (IRT-2p) or the graded response IRT model (IRT-grm), for dichotomous or polytomous items, respectively, either in the normal-ogive or the logistic form; the latter with a scaling constant for approximating the normal distribution function by the logistic one (cf. I. W. Molenaar, 1974). For none of the models applied in the IRT studies does an FA equivalent exist. Of the studies where both FA and IRT are applied, there are two cases that use equivalent models. In one case (Vigneau & Bors, 2005) the Rasch model is applied to dichotomous items. This is equivalent (except for the logistic/normal approximation) to applying a factor model to the matrix of tetrachoric correlations, and restricting the loadings to be equal. In the other case (Wang & Russell, 2005) IRT-grm is applied.

In the remainder of this chapter, we distinguish between FA and IRT in accordance with the authors' terminology. We do so to emphasize the practice as it is presented by the researchers, and because there are only a few studies with equivalent FA and IRT models (six of the 41 studies). Table 2.1 provides an overview of a number of aspects of the studies that are discussed.

*Table 2.1.* Overview of study characteristics.

| | Type of applied analysis | | | |
|---|---|---|---|---|
| Characteristic | FA ($n = 32$) | IRT ($n = 6$) | FA & IRT ($n = 3$) | Total ($n = 41$) |
| Motives provided | | | | |
| no | 24 | | | 24 |
| some | 8 | 5 | 1 | 14 |
| explicit | | 1 | 2 | 3 |
| Type of study | | | | |
| Evaluation | 15 | 2 | 1 | 18 |
| New scale | 8 | 1 | 1 | 10 |
| Translation | 7 | 1 | | 8 |
| DIF | 1 | 1 | 1 | 3 |
| Short form | 1 | 1 | | 2 |
| # Item categories | | | | |
| 2 | 3 | 1 | 1 | 5 |
| 3 | 3 | | 1 | 4 |
| 4 | 5 | 2 | 1 | 8 |
| 5 | 8 | 1 | | 9 |
| 6–8 | 10 | 1 | | 11 |
| varying[a] | | 1 | | 1 |
| no info | 3 | | | 3 |
| # Dimensions | | | | |
| 1 | 1 | 5 | 1 | 7 |
| 2 | 4 | | 1 | 5 |
| 3 | 8 | | | 8 |
| 4 | 4 | | | 4 |
| 5 | 3 | | 1 | 4 |
| 6–15 | 6 | 1 | | 7 |
| varying[b] | 6 | | | 6 |
| Item/factor ratio | | | | |
| median | 7.3 | 18.0 | 12.6 | 7.9 |
| (MAD)[c] | (1.67) | (3.00) | (0.10) | (2.67) |
| Sample size | | | | |
| min. | 118 | 205 | 506 | 118 |
| max. | 9160 | 4306 | 2151 | 9160 |
| median | 577 | 982 | 512 | 553 |
| (MAD) | (368) | (740) | (6) | (351) |
| Respondent/item ratio | | | | |
| median | 17.9 | 41.9 | 20.5 | 18.6 |
| (MAD) | (10.16) | (33.10) | (6.42) | (10.76) |
| Exploratory vs. confirmatory | | | | |
| expl. | 8 | 2 | | 10 |
| conf. | 13 | 4 | 1 | 18 |
| both | 11 | | 2 | 13 |

*Note.* Numbers in the table represent frequencies of studies, except for the row entries min., max., median, and MAD.

[a] A scale consisting of items that differ in the number of categories.

[b] Models with various numbers of dimensions are tested.

[c] MAD: median absolute deviation from the median.

## 2.2 Explicit Model Choice Motives

The first step in gaining information about researchers' motives for applying a certain model is to simply record what investigators themselves say about their motivation. Unfortunately, that is not much.

We distinguish studies where some motives are given for the selection of a (sub)model from studies where no motives are given at all. In addition, we distinguish studies where the model choice is discussed in detail, mentioning both FA and IRT, our primary interest.

In 24 studies (59%), no motives for model choice are given. In 14 studies (34%), some motives are given concerning the choice of the model. In four of the six IRT studies, the benefits of IRT over CTT are described. In some FA studies, the choice of exploratory FA (EFA) versus confirmatory FA (CFA) is defended. Arguments provided in favor of EFA are: "A small items-to-subjects ratio," "not expecting to replicate a factor structure," and "the absence of a previous factor-analytically derived factor structure." Arguments provided in favor of CFA are: "the need for an in-depth analysis of the hypothesized factor structure of the scale," "the possibility of testing a theoretical model," and of "testing competing models by means of comparative fit indices."

In three of the reviewed studies (7%), the model choice is motivated, mentioning both FA and IRT. In two of these (Vigneau & Bors, 2005; Wang & Russell, 2005) both FA and IRT are applied, in the other one (Hong & Wong, 2005) only IRT is applied. Vigneau and Bors (2005) and Hong and Wong (2005) both mention a skewed item distribution as an argument for applying the Rasch model, because it is "insensitive to the shape of the item distributions," whereas in standard (i.e., linear) FA items are assumed to have a multivariate normal distribution if maximum likelihood (ML) estimation is employed.

Because studies where both FA and IRT are applied are of particular interest for the present comparison, we briefly describe each of them. Vigneau and Bors (2005) do not state clearly why they use *both* FA and IRT. It is mentioned that "the IRT model is better suited for the analysis of the dichotomous data." They do not explain why FA is also performed, but perhaps they also want results comparable to previous studies. They perform FA on product-moment, tetrachoric, and corrected phi-correlations ($\phi/\phi_{max}$), which, as they note, have all been used before for the scale under investigation. Based on the FA results of their study, Vigneau and Bors cannot decide whether the data are one- or two-dimensional, whereas the IRT analysis indicates that a unidimensional model does not describe the data well.

Wang and Russell (2005) perform a DIF analysis, i.e., an inspection of whether items function equivalently in different populations of respondents. They describe FA and IRT as "complementary approaches," with IRT better suited for testing equivalence of item parameters (see also Meade & Lautenschlager, 2004), and FA better accommodated for multidimensional model testing. It is remarkable that the CFA is applied on product-moment correlations rather than polychoric correlations, since

the latter would be equivalent to the graded response IRT model that was used. In fact, this relationship is never mentioned in the study.

Remarkably, Clark, Antony, Beck, Swinson, and Steer (2005) apply both FA and IRT without elaborating on the reasons for applying both types of models. They apply IRT at an exploratory stage of scale construction. They determine item discrimination at various levels of the latent variable (LV) by graphically examining item response functions (IRFs). Although not explicitly mentioned, from the references in the article it can be deduced that Ramsay's (2000) TESTGRAF model is used here. Subsequently, the structure of the scale is investigated with a principal component analysis.

Since model choice is not motivated in the majority of the studies, we discuss a number of study characteristics — including aims, some descriptives, and software use — in the next section, in the hope of revealing some implicit motivations.

## 2.3 Characteristics of Data and Applied Models

### 2.3.1 Comparison of Characteristics of FA and IRT Studies

We classify the 41 studies in our analysis into five types, based on their primary aims: evaluation, new scale, translation, DIF, and short form. In 18 of the studies (44%), an existing scale is evaluated. The focus of these studies is usually on the LV structure of the data, examining which items are substantially associated with each other, indicating that they measure the same construct. Another interest of these evaluation studies is to estimate how reliably the items measure the LV.

In 10 of the studies (24%), a new scale is constructed. In these 10 studies researchers often report the process of writing a large number of items, followed by a systematic reduction of the item set in a number of steps, one of which is a psychometric evaluation by means of FA and/or IRT. In eight of the studies (20%), a scale is translated and the psychometric properties of this translated scale are analyzed.

In three studies (7%), a DIF analysis is performed to investigate whether items in the scale are responded to differently by distinct groups of participants.

Finally, in two studies (5%), a short form of an existing scale is constructed and analyzed, with the goal of creating a compact version of a scale consisting of only a small number of items.

From Table 2.1 it can be seen that the type of study is not clearly related to the type of analysis being performed: The relative frequency of the application of FA and IRT is the same for new scale, evaluation, and translation studies. One could argue that in DIF and short-form studies, IRT is applied more often, but these types of studies occurred too infrequently to draw any general conclusions.

The number of item categories varies from two to eight. As is apparent from Table 2.1, five-point scales are most popular for FA studies, but other numbers of categories are also common. IRT studies and dichotomous items are not strongly associated, contrary to what might have been expected. In one study, items vary in

the number of categories, and in two studies no information about item categories is provided.

Thirty-four of our studies (83%) consider multidimensional scales, ranging from 2 to 15 dimensions. In six of these, multiple models with varying numbers of dimensions are tested. Seven studies (17%) consider unidimensional scales. These include one FA study and five of the six IRT studies. It seems that, in practice, IRT is predominantly applied to investigate unidimensional scales.

The ratio of number of items to number of factors varies between 4 and 36 with an overall median of 7.9. In most studies, each factor is represented by 5 to 15 items. This number is well above the recommended minimum of 4 or 5 items per factor for small samples (Marsh & Hau, 1999; Marsh, Hau, Balla, & Grayson, 1998). In the IRT studies, the item/factor ratio is larger than in the FA studies. A confounding factor could be the number of dimensions in the model, as a smaller number of factors with a fixed number of items increases the item/factor ratio.

The sample sizes in the studies vary between 118 and 9160 with an overall median of 553. There are no noticeable differences between FA and IRT studies here, other than that more extreme values are encountered in FA studies, but this could be due to the greater number of FA studies in the sample of articles. Because of the large variation in sample size and the limited number of studies, it is not possible to make any generalizations beyond the reviewed studies about the differences in sample size between FA and IRT studies.

The ratio of number of respondents to number of items varies between 4.6 and 1077 with an overall median of 18.6. In most studies, there are about 20 respondents per item. This number surpasses the ratios of 5 or 10, recommended as lower bounds in the literature (Bentler, 1989; Mueller, 1996; Nunnally, 1978). It should be noted though that such guidelines are too simple. As T. A. Brown (2006, p. 413ff.) notes, many more characteristics of the data and the model should be taken into account to determine a sufficient number of respondents for proper inference. The number of estimated parameters, just to name one, is greater for IRT models than for standard factor models, the former thus requiring more respondents. T. A. Brown suggests choosing a sample size by conducting a power analysis, using either the method proposed by Satorra and Saris (1985) or a Monte Carlo method, in both cases selecting the sample size associated with an 80% likelihood of rejecting a false null hypothesis for the specified model.

In 10 studies (24%), the applied analysis is exploratory; a confirmatory analysis is reported in 18 studies (44%); and in 13 studies (32%), a combination of exploratory and confirmatory analyses is applied. As can be seen from Table 2.1, there are no noticeable differences between FA and IRT studies here.

The software used for scale analysis, as reported in the studies, is shown in Table 2.2. For EFA, either general statistical software (SAS, SPSS, STATVIEW, SYSTAT) is used (four studies) or no information is provided (15 studies), presumably also indicating the use of general software. CFA and IRT are almost always accomplished using specialized software, and information about the software used is almost al-

*Table 2.2.* Overview of software use.

| | Type of applied analysis | | | |
|---|---|---|---|---|
| | FA ($n = 32$) | | IRT | FA & IRT |
| Software | EFA | CFA | ($n = 6$) | ($n = 3$) |
| LISREL | | 12 | | 1 |
| AMOS | | 4 | | |
| EQS | | 2 | | |
| MPLUS | | 2 | | |
| SCA | | 1 | | |
| NOHARM | | | | 1 |
| MSP | | | 2 | |
| RSP | | | | 1 |
| TESTGRAF | | | 1 | |
| MULTILOG | | | | 1 |
| PARSCALE | | | 1 | |
| WINSTEPS | | | 1 | |
| POLY-SIBTEST | | | 1 | |
| EQUATE | | | | 1 |
| DFITPS6 | | | | 1 |
| SAS | 1 | 1 | | 1 |
| SPSS | 1 | | | |
| STATVIEW | 1 | | | |
| SYSTAT | 1 | | | |
| No information | 15 | 2 | | 1 |

*Note.* Numbers in the table represent frequencies of studies. In some studies, multiple software packages are used to estimate different kinds of models.

ways provided. For CFA, LISREL (Jöreskog & Sörbom, 1996) is the most popular, but other packages such as AMOS (Arbuckle, 1995–2006), EQS (Bentler, 1995), and MPLUS (L. K. Muthén & Muthén, 1998–2010) are also used. Researchers who apply IRT use a wide variety of software. In each study, a different program is employed, except for the Mokken scaling program (MSP) which is used twice.

The use of software shows that EFA is more accessible to researchers than CFA or IRT, because it can be carried out with general statistical software. One does not have to obtain and learn how to use a new computer program to do EFA.

It is remarkable that for 17 analyses, no information is provided on the software being used. This is against the policy of the American Psychological Association [APA] (2001, 2010, p. 280/p.210): Although reference entries are not necessary for standard software and programming languages, like SAS and SPSS, the proper name of the software and the version number should always be reported in the text.

### 2.3.2 Types of Studies

Various goals of the analyses are mentioned in the studies. They include the examination of the psychometric properties, the factor structure, and group differences for

a scale. In three studies (all IRT), researchers express their interest in the values of item parameters (such as item difficulty and item discrimination).

When the only analysis is CFA, researchers always (in all 13 studies) mention the goal: testing whether the data fit a hypothesized structure. This objective is, quite remarkably, also mentioned in two of the eight EFA studies. Additionally, in some studies, the choice of specific details of an analysis, e.g., the estimation method or a multilevel approach, is defended.

In Table 2.3 some descriptive statistics are given for the three most prevalent types of studies: new scale development, evaluation of a scale, and translation of a scale. Not surprisingly, researchers constructing a new scale more often apply an exploratory model, whereas researchers evaluating an existing scale more often apply a confirmatory model. However, most, if not all, new scales are developed on a theoretical basis, which would make a confirmatory analysis a reasonable choice. In most studies where an exploratory analysis is reported, researchers do express clear hypotheses about the structure of the data.

It is not clear why many researchers would rather carry out an exploratory analysis than test a hypothesized structure based on substantive knowledge about the items and their mutual relationships, even though the latter approach is far more powerful. Perhaps worries about the presence of secondary factors lure them into an exploratory examination of the factor structure of their data. There is some debate about this issue, with some experts advocating a more liberal application of exploratory techniques (e.g., Bandalos & Finney, 2010). However, we argue that if theory about the underlying structure is available, it is preferable to use that knowledge and suboptimal to neglect it.

*Table 2.3.* Selection of characteristics for the three most prevalent study types.

| | Type of study | | |
|---|---|---|---|
| Characteristic | New scale ($n = 10$) | Evaluation ($n = 18$) | Translation ($n = 8$) |
| Cross-validation | | | |
| no | 4 | 13 | 5 |
| yes | 6 | 5 | 3 |
| Exploratory vs. confirmatory | | | |
| expl. | 5 | 3 | 3 |
| conf. | 2 | 11 | 3 |
| both | 3 | 4 | 2 |
| Sample size | | | |
| median | 462 | 680 | 386 |
| (MAD) | (302) | (426) | (81) |

*Note.* Numbers in the table represent frequencies of studies, except for the row entries median and MAD.

A second reason for applying exploratory rather than confirmatory models could be that CFA (often) requires dedicated software, whereas EFA can be done with general statistical software. The use of specialized software requires an investment — either in time, money, or both — researchers might not be willing to make. This reluctance does not necessarily concern financial costs, since there is a variety of free software available that can handle CFA. We mention OPENMX (Boker et al., 2011; OpenMx Development Team, 2010), and the R packages `sem` (J. Fox, Nie, & Byrnes, 2013; J. Fox, 2006), `lavaan` (Rosseel, 2012), and `lava` (Holst & Budtz-Joergensen, 2012). In addition, free demo versions of both LISREL and MPLUS are available, which are limited with respect to the number of observed variables they can handle.

Sample sizes tend to be somewhat larger for evaluation studies than for other types of studies. This could be due to the fact that an evaluation study is often performed as a secondary analysis of data collected in a large study to investigate properties of the LV the scale is supposed to measure. But again, given the large variation in sample size and the limited number of studies, it is not clear how general the observed pattern is.

## 2.4 Reported Statistical Analyses

### 2.4.1 Descriptive Statistics

Item means are reported in five of the 32 FA studies, in one of the three studies where both FA and IRT are applied, and in four of the six IRT studies. In a fifth IRT study, item difficulty parameters are reported. This difference between FA and IRT applications is to be expected, since one of the focal points of IRT is assessment of item difficulty, of which the item mean is an indicator. In contrast, when the linear factor model is applied, observed variables are often implicitly assumed to be measured as deviations from their means, except when multiple groups are compared (cf. T. A. Brown, 2006, p. 54). In addition, researchers applying IRT might traditionally be more interested in *item* characteristics, whereas researchers who favor FA are more focused on the multidimensional *structure* of the data (cf. Harman, 1968).

In 38 studies (93%), parameters are estimated. The other three studies are nonparametric IRT applications. When reporting parameter estimates, it is recommended to also report corresponding standard error estimates, because they contain information about the variability, hence the reliability of the parameter estimates (e.g., Boomsma, 2000; McDonald & Ho, 2002). This is especially important for studies with small sample sizes or small respondent/item ratios, where standard errors can be relatively large. Nevertheless, standard errors are only provided in four of the 38 studies.

In 27 of the studies (66%), some information is given about the distribution of the estimated latent variable scores in the sample. In 13 of these, information about unweighted sum scores is reported; in one study (Hong & Wong, 2005), latent trait estimate information is provided; and in two studies, the average item scores are given.

In 11 of these studies, it is unclear what kind of latent variable estimate is employed. Information about the distribution usually consists of means and standard deviations (20 studies), but additional distributional properties (e.g., skewness and kurtosis) are also discussed in six studies. In one study (Glutting, Watkins, & Youngstrom, 2005), only the mean is provided.

Correlations between the latent variable estimates are reported in 29 of the 34 multidimensional studies.

### 2.4.2 Model Assumptions

Application of a model is only useful when its assumptions are not violated beyond specific robustness criteria. The use of ML estimation in FA, for instance, requires item responses to have a multivariate normal distribution (Bollen, 1989, p. 107). As another example, the Rasch model in IRT assumes unidimensionality of the item responses.

In 19 of the 32 FA studies, model assumptions are not examined or mentioned at all. In nine FA studies, model assumptions are properly investigated. The distribution of the items is examined and reported upon, and adequate methods are used, such as robust estimators or the use of an appropriate correlation matrix. In four FA studies, model assumptions are considered only marginally: Item distributions are not investigated or described, but a robust estimator is used nevertheless; or researchers describe that they also analyzed their data using robust estimators but do not report these results because the nonrobust analysis gave virtually the same results. The reason for this practice is not entirely clear but perhaps, because of the similarity of results, it is implicitly argued that the use of the robust estimator is not necessary. Moreover, researchers sometimes mention that results from standard methods are more easily comparable with results from other studies. However, this statement is only true when the necessary assumptions are sufficiently satisfied. When the assumptions of a method do not hold, researchers ought to choose an appropriate method, based on the characteristics of the available data only, and report its results.

In four of the six IRT studies, model assumptions are properly examined. The unidimensionality assumption is checked in three studies. Item response functions are examined twice for monotonicity (Sabourin, Valois, & Lussier, 2005; Rivas, Bersabé, & Berrocal, 2005) and once for similarity between observed and estimated functions (Wang & Russell, 2005). In applying the Rasch rating scale model, Hong and Wong (2005) check the assumption of equal spacing of item categories across items. In two IRT applications, assumptions are not given any attention at all.

In the three studies where both FA and IRT are applied, model assumptions are checked properly once, are given some attention once, and are not investigated at all once.

In nine studies (22%), robustness of the estimation method is discussed. Sometimes robust statistics are reported: twice Satorra-Bentler's $\chi^2$ test statistic (Grothe et al., 2005; Shevlin & Adamson, 2005), once an extension of Yuan-Bentler's $T_2^*$ test statistic to multilevel models (Zimprich, Perren, & Hornung, 2005), and once

MPLUS's mean-and-variance adjusted weighted least squares (WLSMV) fit statistic (Leite & Beretvas, 2005). In one study (Toland & De Ayala, 2005), the need for a robust estimator is mentioned, but could not be satisfied and thus not applied, because none was available for the specific method.

In eight of the 26 studies where CFA is applied, the sample covariance matrix is analyzed. The matrix of sample product-moment correlations and the matrix of estimated polychoric correlations are used to estimate the model in four studies each. In two product-moment and in two polychoric correlation studies, ML estimation is employed rather than weighted least squares (WLS), even though WLS is recommended for these matrices, as ML is known to produce erroneous standard error estimates and $\chi^2$-based fit measures when applied to correlation matrices (see, e.g., Cudeck, 1989). In 10 studies, it is unclear what matrix is used for the analysis, which is certainly not a good reporting practice. In one of the 20 studies where EFA is applied, polychoric correlations are analyzed. The other EFA studies either use product-moment correlations or provide no information, which probably also indicates the use of product-moment correlations.

### 2.4.3 Peculiarities

When model assumptions are violated, an analysis may produce unexpected results. Peculiar results may also occur due to other problems, such as model underidentification. Remarkably, none of the studies report peculiarities such as nonconvergence of the estimation procedure or the occurrence of Heywood cases. It is, however, strongly recommended (e.g., Bandalos & Finney, 2010; Boomsma, 2000; De Ayala, 2010) to pay attention to unexpected features of an analysis first, before performing any other evaluation of the results. If no peculiarities are encountered, one should also report that fact.

The lack of reported peculiarities could be explained by a number of reasons. Researchers could fail to notice peculiar outcomes; or would prefer not to report them, because they think that journal editors might not be inclined to accept papers including studies with peculiarities (cf. the "file-drawer" problem; Scargle, 2000).

### 2.4.4 Model Fit and Modification

In 26 of the studies, CFA is part of the analysis. Model fit is formally tested in all of these, except for one (Arrindell et al., 2005), where the multiple group method (Holzinger, 1944) is applied, for which no formal test of model fit exists. The measures most often reported for examining model fit are the root mean squared error of approximation (RMSEA), the goodness-of-fit index (GFI), comparative fit index (CFI), and the Tucker-Lewis index (TLI) or nonnormed fit index (NNFI). Other reported measures are the root mean residuals (RMR) or standardized RMR (SRMR), the normed fit index, the incremental fit index, and the adjusted GFI (AGFI). In addition, the $\chi^2$ test statistic is usually reported with corresponding degrees of freedom and

$p$-value, most often with the annotation that this fit statistic is very sensitive to sample size. It should be noted, however, that the $\chi^2$ test statistic is even more sensitive to nonnormality of the observed variables (Boomsma, 1983). When competing models are compared, reported measures are the $\chi^2$ difference test, Akaike's information criterion (AIC) or a consistent version of the AIC, and the expected cross-validation index.

Factor loadings are reported in 23 of the 26 studies that include a CFA, and are discussed as a criterion for model fit evaluation in six of those studies. The items are then usually required to load significantly on the factor or have a loading higher in absolute value than a criterion value such as 0.40.

Regarding the choice of fit criteria, researchers often refer to one or more statistical publications. In five studies where such criteria are applied, no references are given. In 13 studies, Hu and Bentler (1995, 1998, or 1999) are mentioned. Browne and Cudeck (1993), Hoyle and Panter (1995), Hatcher (1994), and Kline (1998) are referred to in four, three, three, and two studies, respectively. In four studies, other literature is mentioned.

When EFA is conducted, the fit of the model is not formally tested. Instead, researchers use specific criteria to determine the number of factors underlying the items. In 18 of the 21 EFA studies, item factor loadings are examined to determine whether items belong to a factor. Factor loadings greater than 0.30–0.40 in absolute value are usually interpreted as salient. The number of factors is commonly determined by a combination of criteria: examination of a scree plot (15 studies), parallel analysis (10 studies), and/or the eigenvalue $> 1$ criterion (8 studies). In four studies, the percentage of explained variance by a factor is mentioned, without an explicit criterion on how to evaluate it. In only five studies, the interpretability of the factors is mentioned as a criterion. These results might, once again, suggest that many researchers do not use their substantive knowledge about the items to evaluate the structure of the data.

In the IRT studies, formal tests of model fit are never provided. When the Mokken model is applied, Loevinger's $H$-value is reported as an indication of the scale's strength. In three IRT studies, unidimensionality is tested.

The difference between the FA and IRT studies in assessing model fit reflects a difference in traditions. In FA it has become standard practice to report a collection of indices and compare them to cutoff values, a process that has been criticized in the literature (e.g., F. Chen, Curran, Bollen, Kirby, & Paxton, 2008; Marsh, Hau, & Wen, 2004). IRT model fit measures seem to be less well known among researchers, while at the same time claims about the correctness of a model are also more modest than is the case with FA.

In six of the studies where CFA is applied, the model is modified at some point. In three studies, both modification indices and item content are used to modify the model; in one study (Zapf, Skeem, & Golding, 2005), only modification indices are used; in one study (Toland & De Ayala, 2005), only item content is considered; and in one study, the number of factors is adapted after a parallel analysis (Heinitz,

Liepmann, & Felfe, 2005). In the IRT studies, the model is never modified, other than by removing a number of items from the scale.

Some items are discarded in 18 studies (44%). Criteria for item retention include: item content (13 studies); factor loading greater than a certain cutoff value, usually 0.30 or 0.40 in absolute value (11 studies); loading on an additional unintended factor smaller than about 0.20–0.40 in absolute value (eight studies); sufficient item discrimination (two studies); and lack of decrease of the scale's Cronbach's coefficient alpha ($\alpha$; Cronbach, 1951) when the item is excluded (three studies). It should be noted that $\alpha$ typically increases with the number of items in a scale (e.g., Cortina, 1993), and therefore the latter criterion is not very useful. In one study (Beyers, Goossens, Calster, & Duriez, 2005), the criteria used for item retention are not made explicit.

### 2.4.5 Reliability

Reliability refers to the consistency of measurement — that part of a measure that is free of random error (Bollen, 1989, p. 206ff.). The reliability of a scale is estimated in 35 of the studies (85%), usually by computing Cronbach's $\alpha$ for each subscale. It is remarkable that $\alpha$ is still the most commonly used reliability measure, even though other coefficients, like the lower bounds proposed by Guttman (1945), provide greater lower bounds to reliability (e.g., Jackson & Agunwamba, 1977; see also Zinbarg, Revelle, Yovel, & Li, 2005). Moreover, research on the behavior of $\alpha$ (e.g., Cortina, 1993; see also Sijtsma, 2009a, for a historical overview) does not seem to be known. Cortina criticized the practice of comparing $\alpha$ to a cutoff value, like 0.70 or 0.80, without any consideration of context, since the interpretation of $\alpha$ depends on many factors, such as test length and sample homogeneity. Recently, the use of $\alpha$ was criticized and discussed again (Bentler, 2009; S. B. Green & Yang, 2009a, 2009b; Revelle & Zinbarg, 2009; Sijtsma, 2009a, 2009b), with recommendations for alternative reliability estimators and software to employ them. Furthermore, confidence intervals for $\alpha$ (Iacobucci & Duhachek, 2003; Koning & Franses, 2003) are hardly ever reported (three studies, 7%), even though they provide a means of comparing $\alpha$-values from different studies.

In the studies where a nonparametric Mokken IRT analysis is applied, reliability is estimated based on the so-called P(++) matrices. In parametric IRT, one traditionally focuses on item and test information functions, because the measurement error of a scale is a function of the latent variable values. However, in only one of the three parametric IRT applications (Caprara, Steca, Zelli, & Capanna, 2005), are information curves examined to assess reliability. Furthermore, in none of the studies authors report about reliability measures developed within the IRT tradition, such as marginal reliability (B. F. Green, Bock, Humphreys, Linn, & Reckase, 1984) or expected a posteriori (EAP) reliability (Adams, 2005).

In 12 of the 34 studies reporting on a multidimensional scale, a composite reliability measure is provided. In most cases, this is done by computing $\alpha$ for the entire set of items taken together, even though more sophisticated composite reliabil-

ity measures, like weighted $\omega$ (e.g., Bacon, Sauer, & Young, 1995; McDonald, 1970; see also Raykov & Shrout, 2002) are available. Weighted $\omega$ is reported in only one study (Clark et al., 2005). Reliance on the value of $\alpha$ as a lower bound to composite reliability is not justified, as shown by Raykov (1998). He argued that $\alpha$ can be an overestimation (rather than an underestimation) of the composite reliability of a scale when the items have correlated errors.

### 2.4.6 Validity

Cross-validation is performed in 17 studies (41%). In 10 of these, one sample is used to calibrate a proposed structure of the data, and a second, independent sample is used for validation purposes. In seven studies, the sample is randomly split in half to create a calibration and a validation sample. Cross-validation is never applied in the strict sense of imposing the parameter estimates found in the calibration sample on the validation sample, and evaluating the fit (e.g., Camstra & Boomsma, 1992; Cudeck & Browne, 1983). In the examined studies, the common procedure of cross-validation is a CFA on a second sample to test the final model structure that was found by applying an EFA or a CFA on a first sample.

In studies aiming to evaluate a scale, cross-validation is performed remarkably less often than in other types of studies (see Table 2.3). Since an evaluation study is a further examination of an existing scale, researchers might consider their study to be a cross-validation of an earlier study, and argue that an extra cross-validation is unnecessary.

In 24 studies (59%), some external validation of the scale is performed. This is usually accomplished by examining correlations of the scale under investigation with other measures of the construct (convergent validity) or with measures of related but distinct constructs (divergent validity).

### 2.4.7 Expert Coauthor

As a final aspect for comparison, we evaluate some study characteristics in relation to the involvement of researchers with methodological expertise. To this purpose the website or online curriculum vitae of each author was examined. Unfortunately, for six of the studies no information about the authors could be found online. For the remaining 35 studies, we distinguish between (a) studies that are (co)authored by a methodological expert or a psychometrician, (b) studies where the contributions of an expert are acknowledged in an author note, (c) studies where one of the authors shows a research interest in psychometrics or quantitative methods, and (d) studies without any involvement of a methodological expert. Some descriptive statistics regarding this distinction may be found in Table 2.4.

IRT analyses are only performed in studies where one of the authors is a methodological expert, or where an expert's contribution is acknowledged in a note. This also holds for studies where both FA and IRT are applied. It seems that applied researchers

*Table 2.4.* Selection of characteristics for studies with varying degree of involvement of a methodological expert.

| | Methodological expert as coauthor? | | | | |
|---|---|---|---|---|---|
| Characteristic | Yes | Acknowl.[a] | Research interest[b] | No | Unknown[c] |
| Type of analysis | | | | | |
| FA | 11 | | 3 | 12 | 6 |
| IRT | 5 | 1 | | | |
| FA & IRT | 2 | 1 | | | |
| Journal | | | | | |
| EJPA[d] | 4 | | 1 | 2 | 6 |
| EPM[e] | 12 | | 1 | 2 | |
| PA[f] | 2 | 2 | 1 | 8 | |
| Motives | | | | | |
| No | 10 | | | 8 | 6 |
| Some | 5 | 2 | 3 | 4 | |
| Explicit | 3 | | | | |

*Note.* Numbers in the table represent frequencies of studies.

[a] A methodological expert is acknowledged for valuable contributions in a note.

[b] One of the authors has a research interest in quantitative methods.

[c] Information about the authors could not be retrieved from the Internet.

[d] European Journal of Psychological Assessment.

[e] Educational and Psychological Measurement. [f] Psychological Assessment.

without a specific methodological research interest are not sufficiently familiar with IRT to apply it without consulting an expert.

In EPM, most studies are coauthored by a methodological expert. This is not the case for the other two journals. Furthermore, all studies where explicit motives are provided for the choice of methodology — discussing both FA and IRT — are coauthored by a methodological expert. However, it is worth noting that in more than half of the studies coauthored by a methodological expert, no such motives are given. This is rather disappointing and reflects, perhaps, a conflict between authors on the substance of the paper, possibly influenced by limitations on the length of papers accepted by journals.

## 2.5 Summary and Discussion

Although the studies we reviewed are hardly a random sample from all published factor analysis (FA) and item response theory (IRT) applications in psychology and education, we do believe that we can provide a — perhaps limited — impression of the current status of scale construction and evaluation in practice. In fact, we believe this impression to be relatively favorable, since the selected journals offer authors the

opportunity to report on statistically sound research, making it plausible that the reviewed studies are among the methodologically better ones.

We found that FA is applied far more often than IRT. As researchers do not sufficiently explicate their model choice, we are forced to make some educated guesses about possible explanations for this fact. Researchers may not feel obliged to justify their model choice and prefer to use their limited publishing space for different matters, or perhaps they feel uncertain about their choice. In the latter case, the guidelines for applying quantitative methods in the social sciences in a book edited by Hancock and Mueller (2010) might be a useful reference for both authors and reviewers. These guidelines concern model choice as well as reporting practice.

Expectations about the dimensionality of the scale could be a motive to apply FA instead of IRT, even though this is not indicated by the researchers. Researchers' lack of familiarity with software for multidimensional IRT models, or the lack of availability of such software in their research setting, could well be an important reason for IRT's relative unpopularity. FA software is better known: EFA (including principal component analysis) can be conducted using most general statistical packages; for CFA, the LISREL program is most popular and best known. Software for multidimensional IRT models is highly specialized and not very easy to find. Some multidimensional IRT software packages are limited to dichotomous items only, e.g., TESTFACT (Wilson, Wood, & Gibbons, 1984) and MIRTE (Carlson, 1987); polytomous items can be handled by, e.g., MPLUS, CONQUEST (Wu, Adams, & Wilson, 1998), the STATA program `GLLAMM` (Rabe-Hesketh, Skrondal, & Pickles, 2004), and the R package `mcmc` (Martin, Quinn, & Park, 2007).

To our opinion, researchers could take far better advantage of their theoretical knowledge and/or expectations by incorporating their a priori knowledge of the items and scales in the analyses. This should be reflected (a) by a more frequent application of confirmatory techniques, especially in the construction of new scales; and (b) by adding interpretability of factors and content of items to the criteria used for model evaluation.

The issue of applying exploratory versus confirmatory techniques has been discussed by a number of authors (Ferrando & Lorenzo-Seva, 2000; Floyd & Widaman, 1995; Gerbing & Hamilton, 1996; McCrae, Zonderman, Costa, Bond, & Paunonen, 1996). From these studies, although somewhat different in scope, it can be concluded that EFA performs reasonably well at recovering a hypothesized factor structure. Results of CFA and EFA may often be different, with CFA fit measures indicating an unsatisfactory fit of structures uncovered by EFA. The use and interpretation of fit indices, however, is still (or again) under debate (e.g., F. Chen et al., 2008; Marsh et al., 2004; Saris, 2008; Vernon & Eysenck, 2007).

It is troublesome that only in less than half of the studies (46%) model assumptions are investigated to check whether the chosen scaling model is appropriate for the measurement level and distribution of the data, even though well-known guidelines such as those of Wilkinson and the Task Force on Statistical Inference (1999) encourage researchers to do so. Our results are better than those reported by Osborne

(2008), who reviewed 96 articles in the field of educational psychology published in 1998–1999, and found that merely 8.3% reported testing the assumptions of the statistical tests that were used.

Most, if not all, scales in psychological and educational research use ordered categorical items, invalidating the assumption of a linear relation between the items and the latent variable as posed in ordinary FA. Use of this linear model as a pragmatic approach should always be preceded by careful inspection of the distribution of the data. A fair amount of research has dealt with the consequences of applying a linear factor model to polytomous data (e.g., Boomsma, 1983; Coenders, Satorra, & Saris, 1997; Flora & Curran, 2004; Hoogland, 1999; Jöreskog & Moustaki, 2001; Moustaki, Jöreskog, & Mavridis, 2004; B. O. Muthén & Kaplan, 1985, 1992). From these studies it can be concluded that categorical and ordinal data raise no serious problems as long as the distribution of the item variables is approximately normal, illustrating the importance of examining model assumptions. When the distributional assumptions are violated, alternative, robust estimators are proposed which might require specialized software and consultation with methodological experts.

The final question is: What can we learn from the present study, other than "*Most scale researchers use FA, some use IRT, and hardly anyone uses both*"? Most importantly, we learn that far too often models are applied without proper justification. Model assumptions could and should be investigated more frequently. If limited publication space is a bottleneck, authors could consider referring to a website where the results of their analyses would be available for interested fellow researchers. Journal editors may want to encourage such practice by providing journal web space. If expertise is a factor that is lacking, a more frequent collaboration between substantive researchers and statisticians/psychometricians should be encouraged, requiring an active role from both parties. Finally, the education in methodology and statistics for (future) scale developers in the fields of psychology and education might need some reconsideration and reinforcement, again requiring an active role from both substantive researchers and methodological experts.

**Further Research**

As we have seen that assumptions are insufficiently investigated in about half of the studies we have reviewed, some questions arise: What are the consequences of applying a model whose assumptions are violated? What are the advantages of applying a more suitable model? Are there also important disadvantages? And under which specific conditions do these (dis)advantages occur?

Such questions can well be answered by means of Monte Carlo simulation research. In the next chapter we shall give an overview of simulation studies concerning the application of FA and IRT models to ordered categorical data. Findings from this overview of previous research are used to generate specific hypotheses, which are investigated in the simulation study that is presented in subsequent chapters. Finally, and complementary to the Monte Carlo study, in Chapter 7 FA and IRT are applied to

empirical data, much like the data encountered in the studies reviewed in the present chapter.

# Chapter 3

# Previous Research

In this chapter, an overview is given of literature describing Monte Carlo simulation research of factor analysis (FA) and item response theory (IRT). The first section provides a review of simulation studies including both FA and IRT estimation models. Each study is described rather elaborately. To complement this literature overview, simulation research concerned with FA of ordered categorical variables is discussed in an integrated fashion in the second section, while simulation studies on two-parameter IRT models are summarized in the third section. Finally, in the discussion of the reviewed literature, research questions are identified that have not been addressed sufficiently and hypotheses are formulated based on the collected information. These hypotheses will be investigated in the subsequent chapters describing our simulation study.

## 3.1 Simulation Studies Comparing FA and IRT

In the literature, FA and IRT have been compared in a number of Monte Carlo simulation studies, focusing on the application of various models and estimation methods to ordered categorical data. When authors distinguish between *limited-information* FA and *full-information* FA models, we refer to the respective models as FA and IRT, as was explained in Section 1.4.2.

In the next subsections, literature concerning simulation research on the application of FA and IRT to ordered categorical data is summarized in chronological order, by describing the study design, the performance variables, and the main results. As an exception to the chronological order, the last subsection contains some of the findings from an empirical study by Dumenci and Achenbach (2008) that has been included because of its unique contribution of latent variable (LV) score performance variables. The section is concluded by a summary of the literature discussed so far.

In our presentation and discussion of the studies, we use *accuracy* to refer to the degree of conformity of an estimator to the population parameter, as indicated by

the absence of bias of the estimator. When, on average, estimates are too small, an estimator is said to be negatively biased and the parameter or standard error underestimated. Positive bias and overestimation are used analogously. The *precision* of an estimator denotes the closeness of agreement between replications, and is indicated by the low dispersion of the distribution of the estimates.

In Sections 3.2 and 3.3 previous simulation research on the application of FA-only and IRT-only to ordered categorical data is summarized, respectively. The implications of the findings for our Monte Carlo design are given in the final section. Two tables (Tables 3.1a and 3.1b) are presented there, providing a summary of the literature discussed in this chapter.

### 3.1.1 Knol and Berger (1991)

Knol and Berger (1991) compared FA and IRT models for dichotomous items. Data were generated under the normal-ogive two-parameter IRT model (IRT-2p) or the equivalent FA of the estimated tetrachoric correlation matrix (FA-tet), with either one, two, or three LVs.

Various FA estimation methods were included in the comparison, all applied to the matrix of *tetrachoric* correlations. The following FA methods were studied: iterative principal FA-tet (FA-tet-IP; Harman & Jones, 1966), minimum residuals or FA-tet by means of unweighted least squares (FA-tet-ULS; Harman & Jones, 1966), FA-tet by means of generalized least squares (FA-tet-GLS; Jöreskog & Goldberger, 1972), FA-tet by means of maximum likelihood (FA-tet-ML; Jöreskog, 1967), alpha FA-tet (FA-tet-$\alpha$; Kaiser & Caffrey, 1965), and FA-tet by means of an adjusted minimum residuals method (FA-tet-MINR$_{\text{adj}}$; Harman & Jones, 1966; Zegers & Ten Berge, 1983), in which the unique variances are restricted by user-defined lower bounds. In addition, McDonald's (1967) FA by means of a polynomial approximation of the normal-ogive function using unweighted least squares (FA-pa-ULS) was studied. Knol and Berger called this an IRT model, but since only pairwise information on the items is taken into account, it is categorized as an FA model here.

In addition, two IRT models were applied: IRT-2p by means of marginal maximum likelihood with the EM algorithm (IRT-2p-MML; Bock & Aitken, 1981) as implemented in TESTFACT (Wilson et al., 1984) and IRT-2p by means of joint maximum likelihood (IRT-2p-JML), as implemented in MAXLOG (McKinley & Reckase, 1983).

Generated data consisted of 15 dichotomous items loading on one, two, or three LVs, or 30 items loading on three LVs. In each condition LV scores were drawn from a multivariate standard normal distribution. Item difficulty $\beta$ varied between $-2$ and 2. For the unidimensional data, item discrimination $\alpha$ was either 1.00, 1.25, or 1.50; for the multidimensional data, it was either 0 or 1 for each LV. The sample size $n$ was 250, 500, or 1000. The number of replications $R$ was 10. It was restricted to keep computation time manageable for the TESTFACT runs.

Parameters of interest were loadings, thresholds, discriminations, difficulties, and error variance, as well as item response functions (IRFs). Parameter estimation per-

formance was assessed by evaluating the root mean squared error (RMSE) of the estimators, providing a measure of both accuracy and precision.

The main results of the study were:

- Overall, larger samples produced more accurate and precise estimates.
- For the unidimensional data, IRT-2p-MML performed best, the FA methods did reasonably well, and IRT-2p-JML parameter estimation was worst.
- For the multidimensional data, performance of the FA methods was similar to IRT-2p-MML and better than IRT-2p-JML, despite theoretical advantages of the latter two.

Although the IRT models are theoretically more appropriate for the data, the authors concluded that the FA models are preferable because they avoid the numerical problems involved with IRT estimation.

It should be noted that the study was only concerned with dichotomous data. Furthermore, all LV distributions were normal, restricting the generalizability of the results to this type of data. In addition, as the number of replications was very small ($R = 10$), the reliability of the results is limited.

In our study we shall examine whether the same conclusions hold for polytomous items. We shall also examine nonnormal item and LV distributions, and generate a larger number of data sets.

### 3.1.2 Boulet (1996)

Boulet (1996) compared two estimation methods for the two-parameter normal-ogive model: IRT-2p-MML as estimated by the TESTFACT computer program (Mislevy & Bock, 1984) and FA-pa-ULS using NOHARM (Fraser, 1983).

The study was restricted to unidimensional models with dichotomous items, and contained two parallel designs. The first design included three LV distributions: the standard-normal, the $\chi^2_8$ distribution, and the $\chi^2_3$ distribution, resulting in a skewness $\varsigma$ of 0, 1, and 1.63[a], respectively; and excess kurtosis $\kappa$ of 0, 1.5, and 4, respectively. All LV distributions were scaled to have a mean and variance of 0 and 1, respectively, keeping the information/noise ratio constant across LV distributions. Discrimination parameters were 0.5, 1.0, or 1.5. Difficulty parameters were integers between $-2$ and 2. Test length $I$ was 15, 30, 45, or 60 items. Sample size varied between 250 and 4000 such that each test length had three possible sample size/test length ratios, i.e., 16.67, 33.33, and 66.67.

In the second design, sample size was 250, 500, 1000, or 10000, instead of using the sample size/test length ratios. The LV distribution was either standard-normal or moderately skewed here. The remaining variables were identical to those of the first design. This led to 20 additional conditions, since 12 conditions had already been

[a] Boulet reported a skewness of 1.75 for this distribution, which is erroneous, since the skewness of the $\chi^2_3$ distribution is given by $\sqrt{8/3} \approx 1.63$.

covered by the first design and were not sampled again. All data configurations were sampled 100 times.

Performance variables were the signed bias and the RMSE of the parameter estimators. For both difficulty and discrimination parameter estimation, Boulet conducted a repeated-measures analysis of variance (ANOVA) to identify the most influential design factors, only interpreting effects large enough to be considered meaningful quantified as partial $\eta^2 \geq 0.15$.

The main results of the study were:

- IRT-2p-MML produced fewer improper item parameter estimates ($|\hat{\alpha}| > 4.5$ and $|\hat{\beta}| > 4.5$) than FA-pa-ULS.
- With regard to the ANOVA on item discrimination parameter ($\alpha$) estimation, one main effect (estimation method) and five interaction effects (estimation method $\times$ $\alpha$-value, estimation method $\times$ $\beta$-value, estimation method $\times$ LV distribution $\times$ $\beta$-value, estimation method $\times$ $\alpha$-value $\times$ $\beta$-value, estimation method $\times$ $\alpha$-value $\times$ $\beta$-value $\times$ LV distribution) were found.
- When the LV distribution was *normal*, discrimination parameters were recovered more accurately by FA-pa-ULS than IRT-2p-MML. IRT-2p-MML discrimination parameters were negatively biased, and more so for larger discrimination values. FA-pa-ULS discrimination parameter bias was not substantial in case of a normal LV. For both models, estimation of larger discrimination parameters was slightly less accurate for the easiest and most difficult items ($\beta \in \{-2, 2\}$).
- Under mildly *skewed* and extremely skewed LV distributions, IRT-2p-MML discrimination parameter estimators were more accurate than those of FA-pa-ULS, especially for difficult ($\beta \in \{1, 2\}$), highly discriminating ($\alpha = 1.5$) items.
- With regard to the ANOVA on item difficulty parameter ($\beta$) estimation, one main effect (estimation method) and three interaction effects (estimation method $\times$ LV distribution, estimation method $\times$ $\beta$-value, estimation method $\times$ LV distribution $\times$ $\beta$-value) were found.
- For items of moderate difficulty ($\beta \in \{-1, 0, 1\}$), difficulty parameter recovery was reasonably good for both estimation methods and all LV distributions. For items of more extreme difficulty ($\beta \in \{-2, 2\}$), accuracy of difficulty parameter recovery decreased as the LV distribution became more skewed. For nonnormal LV distributions, IRT-2p-MML performed slightly better at recovering extreme difficulty parameters than FA-pa-ULS.
- Neither the sample size, the test length, nor the sample size/test length ratio affected the accuracy of item parameter recovery differentially for FA-pa-ULS and IRT-2p-MML.

As this study included 100 replications, its results are considered moderately reliable. From these results, we can conclude that FA-pa-ULS is the preferred model when model

assumptions are satisfied and discrimination and difficulty parameters are moderate. In case of a nonnormal LV distribution and more extreme parameter values, the additional information taken into account by IRT-2p-MML estimation leads to more accurate and stable parameter estimates than FA-pa-ULS.

### 3.1.3 Finger (2001)

Finger (2001) compared item parameter recovery for the two-parameter normal-ogive model among one IRT and three FA estimation methods. The models under investigation were: IRT-2p-MML, FA-pa-ULS, FA-tet-ULS, and FA of the phi correlation matrix by means of unweighted least squares (FA-phi-ULS), all estimated using his own custom software.

Generated data sets were composed of dichotomous items loading on a standard normal LV. Test length varied between 10 and 50 items, sample size between 250 and 2000. Discrimination parameters were drawn randomly from either a normal $\mathcal{N}(0.75, 0.01)$ or $\mathcal{N}(1.50, 0.04)$ distribution, to result in moderately and higher discriminating item sets, respectively. In each condition, difficulty parameters were drawn from a uniform $\mathcal{U}(-2, 2)$ distribution. For each of the 48 conditions, five replications were generated.

Performance variables were the mean squared error (MSE), the RMSE, and the average signed bias of parameter estimators, as well as the product-moment correlation between population and estimated parameter values, averaged over items. In addition, the MSE of the IRF was calculated by taking the difference between the true and estimated IRF for discretized pieces of the LV continuum (see Finger, 2001, p. 31, for further details). Results were analyzed by conducting multiple univariate ANOVAs on the parameter recovery indices, with the design factors as explanatory variables, interpreting only effects considered meaningful using the criterion of partial $\eta^2 \geq 0.138$.

The conclusions of the study were:

- The accuracy and precision of parameter estimation increased with sample size for all models.
- Test length did not influence the bias of parameter estimators.
- Discrimination parameters were more biased for highly discriminating than for moderately discriminating items. This difference was more pronounced for IRT-2p-MML than for FA-pa-ULS or FA-tet-ULS, and most pronounced for FA-phi-ULS. For IRT-2p-MML and for FA-tet-ULS bias was positive for moderately discriminating items and negative for the highly discriminating items. FA-pa-ULS parameters were all overestimated, whereas FA-phi-ULS parameters were all underestimated.
- Difficulty parameters were unbiased for FA-pa-ULS and FA-tet-ULS. IRT-2p-MML difficulty estimators were negatively biased for highly discriminating items and

unbiased for the less dicriminating item set. FA-phi-ULS difficulty parameter estimators were most biased and overestimated.

- Of all models, FA-phi-ULS parameter estimators were most biased; especially $\alpha$ was recovered very poorly. FA-pa-ULS and IRT-2p by means of marginal maximum likelihood (IRT-2p-MML) performed best.

Finger concluded that both FA-pa-ULS and IRT-2p-MML are appropriate for model estimation. It should be noted that the data studied were limited to dichotomous items and a normal LV distribution. In addition, the small number of replications ($R = 5$) limits the reliability of the results.

### 3.1.4 Tate (2003)

Tate (2003) compared a considerable number of FA and (both parametric and nonparametric) IRT models for dichotomous items: exploratory FA-tet-ULS (EFA-tet-ULS) as implemented in MPLUS, FA-tet by means of mean-and-variance adjusted weighted least squares (FA-tet-WLSMV), also using MPLUS, FA-pa-ULS as implemented in NOHARM, FA-pa-ULS with the addition of a $\chi^2$ dimensionality test estimated by CHIDIM (De Champlain & Tang, 1997), IRT-2p-MML/the three-parameter IRT model by means of marginal maximum likelihood (IRT-3p-MML) using TESTFACT, local item dependence indices for IRT-2p/IRT-3p-LID using IRT-LD (W.-H. Chen, 1993), nonparametric agglomerative hierarchical cluster analysis with a proximity measure based on conditional item pair covariances (IRT-np-HCA) using HCA (Roussos, 1995), nonparametric dimensionality testing (IRT-np-DIM) using DIMTEST (Stout, Douglas, Junker, & Roussos, 1993), and nonparametric simple structure detection (IRT-np-DET) using DETECT (Zhang & Stout, 1999).

Tate applied all models to both empirical and simulated data. Only the design and results of his simulation study are covered here.

Generated data were responses to 60 dichotomous items loading on one, two, or four LVs. The items were divided over the LVs as 60, 30/30, 15/15/15/15, 60/10, or 60/2, where in the latter two cases, 10 and 2 items loaded on both LVs, respectively. The LVs correlated 0.36, 0.60, or 0.90. The LV distribution was logistic, since the item responses were generated with the three-parameter logistic model. The loading parameter was 0.71 for all items, 0.83 for all items, or varied between 0.51 and 0.81. Threshold parameters varied between $-1.41$ and 1.41. The guessing parameter was 0 or 0.20. The sample size was held constant at 2000. These factors were combined to create five unidimensional and 12 multidimensional data configurations, each replicated *once*.

The performance of the models was compared based on the estimation of parameters (for the parametric models only) and the dimensionality assessment of the data (for all models).

The main results from the study were:

- When the population model contained a guessing parameter, FA-tet-ULS and FA-tet-WLSMV failed to correctly recover the dimensionality, and parameter estimates were not accurate. The most extreme condition for either loadings (all $\lambda = 0.83$) or thresholds ($\tau \in \{-1.41, 1.41\}$) deteriorated the results even further.

- The other parametric models also failed to indicate the right number of LVs and to produce accurate parameter estimates when loading or threshold values were extreme. In these cases, the nonparametric models correctly recovered the dimensionality of the data.

- Problematic for nearly all estimation models were the case where two LVs correlated 0.90 (only IRT-3p-MML and IRT-np-HCA performed well) and the case where two of the 60 items loaded on a second LV, representing a violation of local independence (none performed well).

This study is rather limited by the fact that only one replication was generated for each data set. It is nevertheless included in our overview, because it is the only study that included nonparametric IRT models in its design. Results from the study indicate that nonparametric IRT could be a good alternative to parametric models when a guessing parameter is present, loading parameters are as high as 0.80, and item distributions deviate from normality. Whether the latter also holds when guessing is not part of the population model is unclear, because these factors were not investigated independently.

The effect of sample size on model performance was not addressed in this study, as in each condition $n = 2000$. In our study we shall further investigate the possible advantage of nonparametric IRT over parametric estimation models when sample size is small.

### 3.1.5 Kay (2004)

In a setup highly similar to that of Finger (2001), Kay (2004) compared the accuracy and precision of item parameter estimators of one FA and one IRT model: FA-tet-ML, as implemented in PRELIS/LISREL, and IRT-2p-MML, as implemented in BILOG.

The study included unidimensional data with dichotomous items, under both a standard normal $\mathcal{N}(0, 1)$ and a uniform $\mathcal{U}(-3, 3)$ LV distribution. By definition, neither distribution is skewed; the uniform distribution has an excess kurtosis of $\kappa = -1.2$. The sample size was 50, 500, or 1000. Test length was 30, 60, or 100. For each condition, item discrimination was drawn from a normal $\mathcal{N}(0.75, 0.01)$ distribution, and item difficulty from a uniform $\mathcal{U}(-2, 2)$ distribution. For each of the 18 conditions, five replications were generated.

Performance variables were the MSE, the RMSE, and the average signed bias of parameter estimators, as well as the product-moment correlation between population and estimated parameter values, averaged over items. To determine which of the design factors were of importance in affecting the MSE, bias, and correlation, three

univariate ANOVAs were conducted for both the discrimination and difficulty parameter, only interpreting effects considered meaningful using the criterion of partial $\eta^2 \geq 0.138$.

The main conclusions of the study were:

- Larger samples resulted in less bias and more precision of parameter estimators.
- Test length did not affect precision or accuracy of item parameter recovery.
- FA-tet-ML and IRT-2p-MML item difficulty estimators were equally precise and accurate under all conditions.
- Discrimination parameters were estimated more precisely and accurately by IRT-2p-MML than by FA-tet-ML, most markedly in case of the uniform LV distribution and in case of the smallest sample size ($n = 50$).

This study stands out by the smallest sample size included ($n = 50$), for which FA-tet-ML performed notably worse than IRT-2p-MML. Whereas difficulty parameters were recovered equally well by both models, IRT-2p-MML estimation outperformed FA-tet-ML in terms of discrimination parameter estimation, and most markedly for suboptimal conditions, such as a small sample size and a nonnormal LV distribution. Here too, the small number of replications ($R = 5$) limits the reliability of the results.

### 3.1.6 Forero and Maydeu-Olivares (2009) and Forero, Maydeu-Olivares, and Gallardo-Pujol (2009)

Forero and Maydeu-Olivares (2009) compared the graded response model by means of maximum likelihood (IRT-grm-ML) and FA of the estimated polychoric correlation matrix using unweighted least squares (FA-poly-ULS) using the delta and the theta parameterization (see B. O. Muthén, 2006; B. O. Muthén & Asparouhov, 2002). All models were estimated using MPLUS. In an identical setup, Forero, Maydeu-Olivares, and Gallordo-Pujol (2009) compared FA-poly-ULS to FA of the estimated polychoric correlation matrix by means of mean-and-variance adjusted weighted least squares (FA-poly-WLSMV). Therefore, both studies are discussed concurrently here.

The study design was comprised of 324 conditions resulting from crossing six explanatory factors of interest. There were either one or three LVs. The number of items was 9, 21, or 42. The strength of the item-LV relationship was weak (all $\lambda = 0.40$), medium (all $\lambda = 0.60$), or strong (all $\lambda = 0.80$). Items had either two or five response categories. For both response types, three distributional shapes were included: a base type with $p = 0.60$ for the dichotomous items ($\varsigma = 0.40$, $\kappa = 1.17$) and unimodal/symmetric for the five-category items ($\varsigma = 0$, $\kappa = 2.50$); a skewed shape ($\varsigma = 2.00$, $\kappa = 4.84$ for dichotomous items; $\varsigma = 0.98$, $\kappa = 2.80$ for five-category items); and a highly skewed shape ($\varsigma = 2.65$, $\kappa = 8.11$ for dichotomous items; $\varsigma = 1.50$, $\kappa = 4.31$ for five-category items). This resulted in six types of items. The sample size was 200, 500, or 2000. Each cell of the design was replicated 1000 times.

Performance variables were: (a) the convergence rate, or proportion of proper solutions per condition, (b) the relative bias and RMSE of parameter estimators, (c) the relative bias of standard error estimators, and (d) the coverage rate of parameter estimators.

As the theta parameterization of FA-poly-ULS was only included in Forero and Maydeu-Olivares (2009) and the differences with the delta parameterization were minimal, FA-poly-ULS results for both parameterizations are described collectively.

In both studies, the results of the distributional manipulations were presented by distinguishing only between item skewness $\varsigma < 1.50$ and $\varsigma \geq 1.50$. The effects of kurtosis were not discussed separately.

The main results of the study were:

- Convergence rates were better for data sets with a larger sample size and a greater number of items per LV. Skewness had a negative influence on convergence.
- For normal item distributions, parameter and standard error estimators were unbiased with an item/LV ratio $> 7$ for all models under investigation.
- For a combination of suboptimal conditions, loading parameter estimators were positively biased for all models, threshold parameters were biased towards both ends of the latent scale, i.e., negatively signed thresholds were underestimated and positively signed thresholds were overestimated, and loading and threshold standard error estimators were positively biased: (a) small item/LV ratio (3), (b) dichotomous items, (c) small loadings ($\lambda = 0.40$), (d) high item skewness ($\varsigma \geq 1.50$), (e) small sample size ($n = 200$). In these 'harsh' conditions, IRT-grm-ML and FA-poly-WLSMV were more likely to converge than FA-poly-ULS. Parameter estimators by IRT-grm-ML were the most accurate and precise then, followed by FA-poly-ULS and FA-poly-WLSMV.
- The results obtained were quite similar for each of the applied models. Parameter estimation was a little more accurate for FA-poly-ULS than for FA-poly-WLSMV and IRT-grm-ML, whereas standard error estimation was slightly more accurate for IRT-grm-ML, with the FA models overestimating the standard errors to a small degree.
- A disadvantage of IRT-grm-ML compared to FA-poly-ULS was the computation time: IRT-grm-ML model estimation took significantly more time than that of FA-poly-ULS.

The authors concluded that, in preferable conditions, such as approximately normal item distributions, not too few items per LV ($\geq 7$), not too few item categories (5), moderately large loadings ($\geq 0.60$), and not too small a sample size ($\geq 500$), all methods performed equally well in terms of convergence, parameter estimation, and standard error estimation. In that case the choice of estimation model is rather arbitrary. As computation time was shorter for FA-poly-ULS and FA-poly-WLSMV than

for IRT-grm-ML, one might prefer either FA model. However, as computational power accessible to researchers increases rapidly, such considerations become less and less vital.

When a combination of suboptimal conditions is present (skewed, dichotomous item variables loading 0.40 on the LV, few items per LV, small sample size), IRT-grm-ML was most likely to converge and had the least biased parameter and standard error estimators. In these conditions, IRT-grm-ML is clearly the model of choice.

The conclusions of these studies are considered substantial, as the number of replications was large enough ($R = 1000$) to provide reliable results. Furthermore, of all the reviewed studies, our Monte Carlo study is related most closely to these two studies, as will become apparent in this chapter's final section. In our study, we shall further investigate the effects of the distributional shapes of items and LVs, an element not covered in the aforementionded studies.

### 3.1.7 DeMars (2010)

DeMars (2010) compared FA-pa-ULS, FA-tet-WLSMV, IRT-2p by means of robust maximum likelihood (IRT-2p-MLR), and an IRT-2p model based on normal mixtures. FA-pa-ULS was estimated using NOHARM; for the estimation of the other three models, MPLUS was used.

Generated data were dichotomous item responses following the IRT-2p model with a skewed (either positive or negative) or platykurtic (symmetric and thus skewless) LV distribution, the mean and variance of each being set to zero and one, respectively. These distributions were presented graphically, but the kurtosis and skewness were not quantified.

Item discrimination $\alpha$ was 0.30, 0.80, or 1.30. Item difficulty $\beta$ included 15 evenly spaced levels within $[-2, 2]$. The three item discrimination and 15 item difficulty specifications were crossed to create 45 unique items. Sample size was 300 or 5000. The three LV distributions and the two sample sizes resulted in six data conditions, each of which was replicated 1000 times.

To assess the accuracy and precision of the parameter estimators, the bias and the standard deviation of the parameter estimators were examined, respectively. To evaluate the accuracy of the standard error estimators, the ratio of the estimated standard errors to the standard deviation of parameter estimate $\hat{\omega}$, $\hat{se}(\hat{\omega})/sd(\hat{\omega})$, was calculated, and the average was taken over replications.

The main conclusions of the study were:

- FA-pa-ULS and FA-tet-WLSMV results were almost identical. Therefore, the results for both models are discussed collectively, being referred to as FA.

- For each nonnormal LV distribution, discrimination and difficulty parameter estimators were biased for well-discriminating ($\alpha \in \{0.80, 1.30\}$), easy or difficult ($|\beta| > 1.0$) items, but for FA to a greater extent than for IRT-2p-MLR. In more detail, the parameter bias was as follows:

  - For the left-skewed LV distribution, discrimination parameters were positively biased for easy (and thus left-skewed) items, negatively for difficult (and thus right-skewed) items. Difficulty parameters of difficult and easy items were positively biased, regardless of the sign of the difficulty parameter. This effect was larger for larger discrimination values.
  - For the platykurtic LV distribution, discrimination parameters of well-discriminating items were negatively biased for easy and difficult items. This effect was greater for $\alpha = 1.30$ than for $\alpha = 0.80$. Difficulty parameters were biased negatively for easy items and positively for difficult items, and increasingly so for increasingly discriminating items.
  - The parameter bias was smaller for the platykurtic than for the skewed LV distributions, but this might be due to the fact that the kurtosis was less severe than the skewness in terms of deviation from their standard normal values.

- The large sample size ($n = 5000$) combined with low-discriminating items ($\alpha = 0.30$) resulted in accurate standard error estimates for all estimation methods; combined with larger discrimination parameters ($\alpha \in \{0.80, 1.30\}$), standard errors were underestimated. For the smaller sample size ($n = 300$), standard errors for the difficulty parameters were slightly overestimated for low-discriminating items only.

DeMars pointed out that the parameter bias resulting from either FA model could have practical consequences in terms of poor differential item functioning (DIF) estimates. She concluded that IRT-2p-MLR is preferable over FA-tet-WLSMV and FA-pa-ULS when the LV distribution is either skewed or platykurtic.

The study was limited to dichotomous item variables, so the generalizability of the results to polytomous items has yet to be investigated. Furthermore, this study is characterized by the fact that parameter estimates were discussed in discrimination-difficulty combinations. The bias of difficulty parameter estimators was larger when item discrimination was larger. This is understandable, when considering that (a) both $\alpha$ and $\beta$ depend on $\lambda$, and (b) the bias of $\lambda$ increases with its absolute value. Hence, both discrimination and difficulty parameter bias are larger for larger loading values.

In addition, the accuracy of discrimination parameter estimators depended largely on the difficulty of the item under investigation. The increase in discrimination parameter bias for items of more extreme difficulty could be due to the fact that, for dichotomous items, extreme item difficulty is expressed in extreme skewness of the item distribution.

In our study, we shall investigate whether these results also hold for polytomous items. In addition to discrimination and difficulty parameters, we shall examine loading and threshold parameters, which are less interdependent than discrimination and difficulty parameters, thus facilitating the interpretability of results.

### 3.1.8 Finch (2010, 2011)

Finch (2010) compared FA-pa-ULS as implemented in NOHARM, FA-tet-WLSMV as implemented in MPLUS, and IRT-2p-MML using BILOGMG. In a subsequent study, Finch (2011) extended his 2010 design, including only FA-pa-ULS and IRT-2p-MML, however. Since the two studies share a common setup, both are discussed simultaneously.

In the *first* study, data were simulated with two LVs underlying 15, 30, or 60 dichotomous item variables, where half of the items loaded on each LV. The LV distributions were either standard normal or skewed ($\varsigma = -1.5$, $\kappa = 3.0$). The correlation between the LVs was 0, 0.3, 0.5, or 0.8. Item discrimination parameters were drawn from a normal $\mathcal{N}(0.97, 0.32)$ distribution truncated at 0.37 and 2.02, based on empirical findings. Item difficulty parameters were drawn from a standard normal $\mathcal{N}(0, 1)$ distribution. Sample size was 250, 500, 1000, or 2000.

In the *second* study, Finch focused on data *not* exhibiting simple structure when attempting to fit a simple-structure model. For this purpose, data were simulated with two LVs and 30 dichotomous item variables. Twenty-six items loaded on one LV (13 each) according to the loading specification of the first study. In addition, four items loaded on both LVs. Two factor structure conditions were introduced: In one condition, the cross-loadings were half the size of the main loading; in the other condition, the cross-loadings equaled the main loading sizes. In every condition, a simple structure model was tested confirmatively, so the cross-loadings were deviations from the hypothesized model. The other design parameters were identical to those of the first study. Each data condition was replicated 1000 times in the first and 500 times in the second study.

Data were also generated with a pseudo-guessing parameter, resulting in a three-parameter IRT model. These results are discussed only briefly here, because the scope of this dissertation is limited to two-parameter IRT or common FA models.

Because the IRT-2p-MML model allows for only one LV, two unidimensional IRT-2p-MML models were estimated — with half of the items loading on each LV — for every data configuration. As the FA models were suitable for estimation of the multidimensional model, they were applied as such.

Performance variables were the parameter bias, size of the corresponding standard errors, and RMSE of the discrimination and difficulty parameter estimators. In addition, an ANOVA was conducted to determine which of the design factors explained most of the variance of the RMSE of the discrimination and difficulty parameter estimators.

The main results of both studies were:

- In case of a skewed LV, difficulty parameter estimators were biased, especially for FA-pa-ULS; standard error estimates of both discrimination and difficulty parameters were larger compared to the normal LV condition for all models, most notably for FA-tet-WLSMV discrimination parameter estimators.

- When a guessing parameter was part of the population model, FA-tet-WLSMV discrimination and difficulty parameters were severely under- and overesti-

mated, respectively. In this case, IRT-2p-MML and FA-pa-ULS also exhibited larger parameter estimation bias, but to a far lesser extent.

- Sample size and test length did not significantly affect the accuracy of any of the discrimination or difficulty parameter estimators under investigation. The standard errors of these estimators decreased with increasing sample size for FA-pa-ULS and FA-tet-WLSMV.
- Discrimination and difficulty parameter estimators of items loading on multiple LVs were more biased than those of simple structure items. This difference was far more pronounced for IRT-2p-MML than for FA-pa-ULS. Generally, discrimination parameters were overestimated, whereas difficulty parameters were underestimated.
- FA-pa-ULS discrimination parameter estimators were more biased for items loading unequally on two LVs (one dominant LV) than for items loading equally on two LVs. For FA-pa-ULS difficulty parameter estimation, the result was reversed.
- The larger the correlation between the LVs, the larger the bias of parameter estimators and the larger their standard errors.

Finch concluded that repeatedly applying a unidimensional model for each dimension, as was done with IRT-2p-MML, was a viable alternative to confirmatively employing a multidimensional model with a correctly specified item-LV structure and a large test length (60 items). However, ignoring the multidimensional structure of the data completely led to biased item discrimination and difficulty parameter estimators. Furthermore, when items were the product of multiple LVs, i.e., when simple structure was not present in the data, parameter estimators were generally biased. When the LV distribution was skewed, the direction of the difficulty parameter bias equaled the direction of the skewness.

### 3.1.9 Maydeu-Olivares, Cai, and Hernández (2011)

Maydeu-Olivares, Cai, and Hernández (2011) compared the fit of FA of the sample covariance matrix by means of maximum likelihood (FA-lin-ML) and IRT-grm-ML, using personal software routines for both.

They argued that in order to compare the fit of FA-lin-ML and IRT-grm-ML models, one cannot use the standard $\chi^2$ test statistic, since it assumes that the response variables are normally distributed, which is, by default, not the case for ordered categorical data. The authors compared the performance of three fit statistics: (a) Browne's (1984) $\chi^2_B$ statistic, which is known to overreject the model for finite samples and empirically relevant test lengths; (b) Yuan and Bentler's (1997) $\chi^2_{YB}$, which is a correction to the $\chi^2_B$; and (c) Yuan and Bentler's (1999) $F_{YB}$, which is an $F$-distributed correction to the $\chi^2_B$.

In their Monte Carlo design, data were 10 or 20 continuous, dichotomous, or five-category items loading on one normal LV. However, FA-lin-ML was applied to the

data consisting of continuous items, whereas IRT-grm-ML was applied to the ordered categorical data. Loading parameters varied between 0.4 and 0.8. For the categorical data, difficulty parameters were either 0 or 0.97. For the continuous data, item distributions were multivariate normal with zero intercepts. Sample size equaled 500, 1000, 2000, or 5000. Each data configuration was replicated 1000 times.

The rejection rates of each of the three fit statistics at a 5% significance level were evaluated for each model. The authors found that the statistics performed similarly for FA-lin-ML and IRT-grm-ML. The $\chi^2_{YB}$ test statistic was found to perform best, with useful rejection rates even for the smallest sample size ($n = 500$) and the largest model ($I = 20$).

The authors went on to investigate the performance of the fit statistics by examining the difference in rejection rates between FA-lin-ML and IRT-grm-ML for ordered categorical data. They concluded that none of the investigated indices are very useful for distinguishing between both models, as the fit of IRT-grm-ML was only marginally better than the fit of FA-lin-ML, with one exception: the $\chi^2_{YB}$ statistic indicated a better fit of IRT-grm-ML than FA-lin-ML to data sets of 20 dichotomous items and $n = 5000$.

### 3.1.10 Dumenci and Achenbach (2008)

The differences between LV scores resulting from FA and IRT models were addressed by Dumenci and Achenbach (2008), who used *empirical* data to compare LV score estimation of, among others, FA-lin-ML, FA-poly-WLSMV, IRT-grm-EAP[b], and simple sum scores.

The empirical data under investigation were three psychopathology scales, consisting of 11, 18, and 18 three-category items, which were all moderately to highly right-skewed. Presumably, the LVs underlying these items were also right-skewed.

In their study, two groups of methods emerged based on their LV score estimate properties: methods that did not take into account the ordinal properties of the data (FA-lin-ML and simple sum scores) and methods that did (IRT-grm-EAP and FA-poly-WLSMV). Within these groups, differences between LV score estimators were practically nonexistent. Visual inspection of the relations between the LV scores as estimated by the two groups, using scatterplots, revealed strong nonlinear associations. Fitting a linear curve on the bivariate relations explained about 87% of the variance in the LV scores; applying a quadratic and cubic function in the fitting procedure increased the explained variance to 96% and 98%, respectively. Transformation of the LV scores to rank orders showed that the ordering of respondents based on the LV score estimates was similar for all methods.

Dumenci and Achenbach concluded that FA-lin-ML LV score estimators and simple sum scores were biased towards the center on both ends of the LV distribution, which could have serious practical consequences, since in diagnostic settings or other clinical practices one is often particularly interested in respondents having exceptionally high

[b]The graded response model by means of Expected a posteriori estimation.

or low LV scores. Consequently, they recommended the use of the graded response IRT model (IRT-grm) or FA-poly-WLSMV when using ordered categorical items for estimating LV scores. They also suggested investigating this matter in a simulation study.

In our Monte Carlo study, we shall evaluate LV score estimates as a performance variable, as will be discussed in the final section of this chapter.

### 3.1.11 Summary

In this subsection we summarize the studies discussed so far, thematically ordered. Since DeMars (2010) found almost identical results for FA-pa-ULS and FA-tet-WLSMV, we tentatively generalize their FA-pa results to FA-tet.

When the LV distribution was normal, discrimination parameters were recovered about equally well by FA-tet/FA-pa and IRT-2p (Boulet, 1996). The fact that Kay (2004) and Knol and Berger (1991) found somewhat better results for IRT-2p-MML, whereas Finger (2001) found FA-pa-ULS parameter estimators to be slightly more accurate, is probably due to these studies' limited number of replications (5, 10, and 5, respectively). For polytomous models, Forero and Maydeu-Olivares' (2009) results were consistent with Boulet's.

Under mildly skewed and extremely skewed LV distributions, parameter estimation by IRT-2p-MML was more accurate and precise than by FA-pa-ULS (Boulet, 1996; Finch, 2010). This was even more so when the signs of the item skewness and the LV skewness were opposite (Boulet, 1996). For a uniform LV distribution, IRT-2p-MML discrimination parameter estimators were more precise and accurate than those of FA-tet-ML (Kay, 2004).

For a more skewed LV distribution and more extreme item difficulty, i.e., skewed item distributions, the accuracy of FA-pa-ULS and IRT-2p-MML difficulty parameter estimators decreased, regardless of the direction of item skewness (Boulet, 1996).

For well-discriminating ($\alpha \in \{0.8, 1.3\}$), skewed items, both discrimination parameters and difficulty parameters were biased, to a greater extent for FA-tet-WLSMV than for IRT-2p-MLR (DeMars, 2010). Discrimination parameters were underestimated when the sign of LV skewness and item skewness matched and were overestimated when it did not match. Difficulty parameters were overestimated for a left-skewed LV and underestimated for a right-skewed LV, regardless the sign of item skewness. For more uniform items (middle difficulty) difficulty parameters were underestimated for left-skewed and overestimated for right-skewed LVs.

Forero and Maydeu-Olivares (2009) found that IRT-grm, and FA-poly parameter and standard error estimators were biased for (a) a small item/LV ratio (3), (b) dichotomous items, (c) small loadings ($\lambda = 0.40$), (d) high item skewness ($\varsigma \geq 1.5$), and (e) small sample size ($n = 200$). For a very small sample size ($n = 50$) FA-tet-ML parameter estimators were far less accurate and precise than those of IRT-2p-MML (Kay, 2004).

Higher loading/discrimination values led to more bias when the item distribution was determined by a nonnormal LV distribution (Boulet, 1996; DeMars, 2010).

When the LV distribution was normal and item nonnormality was determined by threshold manipulation, smaller loading/discrimination values worsened parameter and standard error estimation (Forero & Maydeu-Olivares, 2009).

We conclude that, in absence of model violations and in preferable conditions (items/LV ratio $\geq 7$, item categories $\geq 5$, loadings $\geq 0.60$, $n \geq 500$), FA-poly and IRT-grm performed equally well in terms of convergence, parameter estimation, and standard error estimation (Forero & Maydeu-Olivares, 2009). Under suboptimal conditions (skewed, dichotomous item variables, loading 0.40 on a normal LV, or loading strongly on a nonnormal LV; few items per LV; small sample size), IRT-grm seems to outperform FA-poly (Boulet, 1996; DeMars, 2010; Forero & Maydeu-Olivares, 2009; Kay, 2004).

Although nonparametric IRT (NIRT) has not received much attention in simulation research, results from Tate (2003) indicate that NIRT could be a good alternative to parametric models when a guessing parameter is present, loading parameters are as high as 0.80 and item distributions deviate from normality. Whether the latter also holds for the nonparametric Mokken IRT model (IRT-mok), when guessing is not part of the population model, is one of the topics of our simulation study.

## 3.2 Results from FA-only Simulation Studies

Robustness questions have been addressed frequently in FA research. In the current section, a summary is provided of the main conclusions resulting from simulation studies concerned with FA of ordered categorical data.

In the literature, a number of studies have been concerned with the comparison of various FA models applied to categorical data. In these studies, the correlation matrix under investigation varied as well as the estimation method. In the following, the results for FA of the sample covariance matrix (FA-lin) and FA of the estimated polychoric correlation matrix (FA-poly) are summarized.

### 3.2.1 Findings for Linear Factor Analysis

FA-lin-ML *parameter* estimators were negatively biased for approximately normal five-category items with a sample size up to 700 (Babakus, Ferguson, & Jöreskog, 1987; Beauducel & Herzberg, 2006; Coenders et al., 1997; DiStefano, 2002; Dolan, 1994; Rhemtulla, Brosseau-Liard, & Savalei, 2012; Trierweiler, 2009). B. O. Muthén and Kaplan (1985) reported no parameter bias for $n = 1000$, whereas Beauducel and Herzberg (2006) and Trierweiler (2009) found loading parameter estimators to be negatively biased for that sample size.

Parameter bias was worse for skewed item variables (Babakus et al., 1987; Boomsma, 1983; Coenders et al., 1997; DiStefano, 2002; Rhemtulla et al., 2012). As skewness (asymmetry) is always accompanied by kurtosis, whereas the reverse does not necessarily hold, B. O. Muthén and Kaplan (1985) tried to separate the two by including a symmetric kurtotic item distribution in their design. Because parameter

estimators for items following this distribution were not found to be more biased than approximately normal items, it was concluded that skewness rather than kurtosis was the main cause of bias.

Under the condition of nonnormal underlying item variables and evenly spaced thresholds, Coenders et al. (1997) found maximum likelihood (ML) loading estimators to be less biased than weighted least squares (WLS) estimators for FA-poly when $n = 1000$. Rhemtulla et al. (2012) found a similar amount of parameter bias for FA-lin-ML and FA-poly-WLSMV under the condition of a positively skewed LV distribution and evenly spaced thresholds for sample sizes between 100 and 600. Bias was negative for FA-lin-ML and positive for FA-poly-WLSMV, however.

The literature is inconclusive with regard to FA-lin-ML *standard error* estimation. Whereas Babakus et al. (1987) found FA-lin-ML standard errors to be overestimated for sample sizes up to 500 for both normal and nonnormal item distributions, DiStefano (2002) found substantial negative bias only for skewed item variables. B. O. Muthén and Kaplan (1985) reported negative standard error estimator bias only for highly skewed or kurtotic items and item correlations larger than 0.5. Rhemtulla et al. (2012) reported consistent underestimation of robust FA-lin-ML standard errors of loading parameters, particularly for small sample sizes ($n \in \{100, 150\}$).

For practical purposes, FA-lin-ML *fit indices* were satisfactory when the model was specified correctly and item distributions were normal (Beauducel & Herzberg, 2006; Rhemtulla et al., 2012; Trierweiler, 2009), but for skewed items the chi-square, goodness-of-fit index (GFI), and nonnormed fit index (NNFI) tended to indicate underfit (Boomsma, 1983; DiStefano, 2002; B. O. Muthén & Kaplan, 1985; Rhemtulla et al., 2012). The standardized root mean residuals (SRMR) and root mean squared error of approximation (RMSEA), however, seem to be quite robust against nonnormal item distributions (DiStefano, 2002).

### 3.2.2 Findings for Polychoric Factor Analysis

In the literature, the performance of FA-poly is reported on as estimated by ML, robust ML (MLR), generalized least squares (GLS), unweighted least squares (ULS), WLS, diagonally WLS (DWLS), mean-and-variance adjusted WLS (WLSMV), and a categorical variable methodology (CVM) (B. O. Muthén, 1984).

For item distributions with skewness up to 1.5, FA-poly-ML parameter and standard error estimators were found to be unbiased (Babakus et al., 1987; Dolan, 1994; Yang-Wallentin, Jöreskog, & Luo, 2010). Babakus et al. and Dolan found that FA-poly-ML $\chi^2$ values were overestimated, especially for the conditions with skewed item variables and sample sizes ranging from 200 to 500. Yang-Wallentin et al. found FA-poly-ML $\chi^2$ values to be only slightly overestimated for sample sizes ranging from 400 to 1600, for both approximately normal and skewed items. The adjusted GFI and root mean residuals (RMR) fit indices were rather poor, especially with $n = 100$ or with nonnormal item distributions (Babakus et al., 1987).

Loading parameters were overestimated considerably by FA-poly-GLS (Dolan, 1994). FA-poly-ULS parameters were unbiased for correctly specified models for both nor-

mal and skewed items, and two-, five-, or seven-category items (Yang-Wallentin et al., 2010). For dichotomous items, Parry and McArdle (1991) found that item skewness led to loading parameter bias for FA-tet-ULS. FA-poly-ULS standard error estimators were unbiased and as good as FA-poly-DWLS standard error estimators (Yang-Wallentin et al., 2010). FA-poly-GLS standard errors were unbiased for both normal and skewed item variables (Dolan, 1994). Yang-Wallentin et al. found ULS $\chi^2$-estimators to be unbiased and correctly indicative of model misspecification when it applied.

Potthast (1993) investigated the performance of FA-poly-CVM as implemented in LISCOMP (B. O. Muthén, 1987) under conditions of nonnormal item variables. She found loading parameter estimators to be unbiased. Coenders et al. (1997) also found practically no parameter bias in case of skewed LV or item distributions. Standard errors were unbiased for unidimensional and two-dimensional data with a large sample size ($n = 1000$). Remarkably, standard error estimators for negative kurtotic items were less biased than those of approximately normal items (Potthast, 1993). For one- and two-dimensional data, $\chi^2$ estimates were close to their expected values; for three- and four-dimensional data, however, they were severely overestimated. The overestimation increased with increasing kurtosis of the items and decreasing sample size.

Parameters estimated by FA-poly-WLS were unbiased for approximately normal items starting from a sample size of 500, when the model was correctly specified (Coenders et al., 1997; DiStefano, 2002; Dolan, 1994; Flora & Curran, 2004; Yang-Wallentin et al., 2010). For skewed items, parameter estimators were only biased when the skewness was due to skewness of the LVs; when item skewness was the result of manipulation of the threshold values, while retaining a normal LV, parameter estimators were unbiased (Coenders et al., 1997; DiStefano, 2002; Flora & Curran, 2004; Yang-Wallentin et al., 2010). Compared to FA-poly-ML, FA-poly-ULS, and FA-poly-DWLS, FA-poly-WLS performed worst, structurally overestimating loading parameters (Flora & Curran, 2004; Trierweiler, 2009; Yang-Wallentin et al., 2010).

FA-poly-WLS standard error estimators have consistently been found to be negatively biased for sample sizes up to 1600 (Coenders et al., 1997; DiStefano, 2002; Dolan, 1994; Flora & Curran, 2004; Trierweiler, 2009; Yang-Wallentin et al., 2010).

FA-poly-WLS $\chi^2$ values were overestimated for sample sizes ranging from 100 to 1600, inflating the model rejection rate (DiStefano, 2002; Flora & Curran, 2004; Yang-Wallentin et al., 2010). Contradictory to this result, Holgado-Tello, Chacón-Moscoso, Barbero-García, and Vila-Abad (2010) found FA-poly-WLS $\chi^2$ values to be underestimated for the sample size of 1000. The SRMR fit index was inflated for the medium sample size of 350 (DiStefano, 2002). The GFI and the NNFI performed well in not rejecting a correctly specified model, even when items were highly skewed; and the RMSEA correctly indicated a good fit for sample sizes starting from 350 (DiStefano, 2002; Holgado-Tello et al., 2010); for smaller sample sizes, the RMSEA was found to be inflated (Trierweiler, 2009).

Loading parameters, as estimated by FA-poly-DWLS/FA-poly-WLSMV, were not substantially biased for as few as five normal and skew-normal items (skewness up to 1.53) loading on a single LV, for sample sizes as small as 100 (Beauducel & Herzberg, 2006; Flora & Curran, 2004; Forero et al., 2009; Rhemtulla et al., 2012; Trierweiler, 2009). With a positively skewed LV, Rhemtulla et al. found FA-poly-WLSMV parameters to be overestimated, but when item distributions were manipulated by setting threshold values, parameter estimation was not affected.

FA-poly-WLSMV loading standard error estimators were found to be negatively biased for small sample sizes for normal and skew-normal LV distributions, and increasingly so with greater skewness (Flora & Curran, 2004; Rhemtulla et al., 2012; Trierweiler, 2009). With $n \geq 500$, these estimators were not substantially biased under normal and skewed/kurtotic LV conditions (Flora & Curran, 2004; Trierweiler, 2009).

Flora and Curran (2004) found FA-poly-WLSMV $\chi^2$ values to be overestimated for sample sizes up to 1000, and bias to be smaller for a 5-item than for a 10-item scale. Rhemtulla et al. (2012), focusing on the rejection rate, found the FA-poly-WLSMV $\chi^2$ to perform very well. Trierweiler (2009), simulating data with either three or six items per LV, found hardly any FA-poly-DWLS $\chi^2$-estimator bias (using LISREL) for sample sizes up to 1000. The only case where she found the $\chi^2$ values to be considerably overestimated was when data consisted of six items per LV with moderately positive and negative skewness ($|\varsigma| = 1.53$) for sample sizes ranging from 100 to 1000. With a correct model specification, the comparative fit index (CFI), Tucker-Lewis index (TLI), RMSEA, and SRMR all performed well (Beauducel & Herzberg, 2006; Trierweiler, 2009).

Trierweiler (2009) found that FA-poly-MLR loading parameter estimators were unbiased for normal and skew-normal item distributions (skewness up to 1.53). FA-poly-MLR standard error estimators were negatively biased under two conditions: (a) when items were skewed and sample size was small ($n = 100$); and (b) when items were skewed, sample size was small to moderate ($n \leq 200$), and the item/LV ratio was small (3). Both for approximately normal and for skewed items, $\chi^2$ values were unbiased. Consistent with her FA-poly-DWLS results, for FA-poly-MLR Trierweiler also found substantial overestimation of $\chi^2$ values, only when data consisted of six items per LV with moderate positive or negative skewness ($\varsigma = 1.53$) for all sample sizes (100–1000). The CFI and RMSEA fit measures correctly indicated a close model-data fit.

## 3.3 Results from IRT-only Simulation Studies

Two-parameter IRT model has been investigated by means of Monte Carlo simulation in a number of studies (Drasgow, 1989; Maydeu-Olivares, Drasgow, & Mead, 1994; Parshall, Kromrey, Chason, & Yi, 1997; Stone, 1992; Tuerlinckx & De Boeck, 2001). As the focus and setup of these studies vary considerably, it is not possible to follow the same line of presentation as was chosen for the FA-only studies. Instead, the

results of each of the IRT-only studies will be summarized separately in a brief outline. Unfortunately, simulation research on NIRT directly relevant to our studies is not available. Therefore, in the final subsection, the available research concerning NIRT is addressed only by reference.

### 3.3.1 Drasgow (1989)

Drasgow (1989) compared the performance of IRT-2p-MML and IRT-2p-JML using his own FORTRAN code. He focused on the accuracy of discrimination and difficulty parameter and standard error estimators for models with dichotomous items loading on a single LV, under varying conditions of test length (5–25), item discrimination (0.4–1.8), item difficulty (−1.5–2.5), and sample size (200–1000). Each condition was replicated 10 times. It was found that IRT-2p-MML estimation was generally more accurate than IRT-2p-JML estimation of both parameters and standard errors. For items of moderate discrimination and difficulty, IRT-2p-MML parameter and standard error estimates were accurate for a scale of five items and a sample size as small as 200. Discrimination parameter estimators for items of extreme discrimination were severely overestimated when test length was small ($I = 5$), regardless of the sample size. IRT-2p-MML difficulty parameter estimator bias (the sign of which is not clear from the article), however, did decrease with increasing sample size and was only unacceptably large for a scale of five items and the smallest sample size ($n = 200$).

### 3.3.2 Stone (1992)

Stone (1992) investigated the robustness of IRT-2p-MML estimators as implemented in MULTILOG. He evaluated the bias and RMSE of discrimination and difficulty parameters, the test response function, and LV scores, under varying conditions of test length (10–40), item discrimination (0.72–3.00), item difficulty (−2.2–2.4), and sample size (250–1000). Each data configuration was replicated 100 times. Items were dichotomous and loaded on a single LV that was standard normal, positively skewed ($\varsigma = 0.75$, $\kappa = 0.0$), or symmetric/platykurtic ($\varsigma = 0.0$, $\kappa = -1.0$).

Stone found that discrimination parameter bias was small but negative for the normal LV; the bias as well as the RMSE increased as the population parameter value increased, and decreased with increasing sample size. Discrimination parameters were overestimated for items loading on the skewed or platykurtic LV; this bias and RMSE decreased as the number of items increased from 10 to 40, and was larger for the skewed LV than for the platykurtic LV.

Difficulty parameters demonstrated a similar pattern: They were biased for the nonnormal LV distributions for moderate to large difficulty values $|\beta| > 1$, and the absolute value of bias was larger for the skewed than for the platykurtic LV distribution. It was negative when population values were negative and positive when population values were positive, regardless of the LV distribution.

The bias of LV score estimators was not much affected by any of the explanatory factors under investigation. The RMSE decreased as test length increased. Absolute

bias and RMSE increased as the absolute true LV values increased. Scores in the tails of the distribution were estimated to be closer to the middle.

### 3.3.3 Maydeu-Olivares et al. (1994)

Maydeu-Olivares et al. (1994) compared the fit of IRT-grm-ML and the two-parameter partial credit IRT model by means of ML (IRT-pc2-ML) as estimated by MULTILOG. They investigated whether it was possible to distinguish between these two models under varying conditions of test length (5–25) and sample size (250–1000). The population model was either IRT-grm or IRT-pc2. Data were unidimensional and the number of item response categories was held constant at five. For each condition three samples of data were generated. The largest effect was found for test length: The more items, the better the two estimation models could be distinguished, i.e., for a longer test, the model fit of the true model was much better compared to the other model, whereas for a shorter test this difference was much less pronounced. Model distinction also ameliorated with a larger sample size, and this effect was larger for longer tests. Overall, results for IRT-grm-ML and IRT-pc2-ML were very similar, which led the authors to conclude that either model would be equally appropriate in practical applications of scale analysis.

### 3.3.4 Parshall et al. (1997)

Parshall et al. (1997) compared the one-, two-, and three-parameter IRT models (IRT-1p, IRT-2p, and IRT-3p), and modified versions of the latter two (restricted discrimination parameter for IRT-2p and IRT-3p, and a restricted discrimination as well as a constant guessing parameter for IRT-3p). BILOG was used for estimation, by means of marginal ML using EM. Data were generated as 80 dichotomous items from a six-dimensional normal IRT-3p population model, using NOHARM. The authors used multidimensional data to account for the multiple abilities that real-life respondents typically use when taking a test. Sample size varied between 100 and 1000. Each data configuration was replicated 100 times.

The accuracy of item parameters per se could not be evaluated, because the data generation model did not match any of the estimation models. Instead, the authors compared the estimated *response patterns* to the true response patterns. In addition, the standard deviations of the parameter estimates were examined to determine the precision of the parameter estimators. The fit of each model was assessed using cross-validation samples: Twenty new samples of $n = 100$ were generated. The fit of the estimated item parameters to these new samples of data was assessed for each replication.

It was found that the models with more free parameters (or fewer constraints) resulted in a better fit to the data, but also in lower levels of precision of parameter estimators as evidenced by larger standard deviations. As for the cross-validation results, the more free parameters — and thus the better the calibrated fit — the worse the cross-validated fit, which is indicative of overfitting or capitalizing on chance. In

addition, fit was worst at the largest sample size, also suggesting overfitting to the idiosyncrasies of the data.

### 3.3.5 Tuerlinckx and DeBoeck (2001)

Tuerlinckx and De Boeck (2001) investigated the violation of the assumption of local independence of items for IRT-2p-MML as implemented in MULTILOG, under varying conditions of item interdependencies (positive, negative interactions of various strengths, as modeled by Hoskens and De Boeck's (1997) constant interaction model), test length (6–20), item difficulty (−2–2), and sample size (500–1000). Results were discussed in terms of mean estimated discrimination parameters. For the six-item conditions, the proportion of replications where the likelihood ratio test of fit rejected the model was also reported. This test statistic could not be calculated for scales longer than six items, as the contingency table would then contain too many cells compared to the number of respondents.

The authors found that a positive interaction between items, i.e., positively associated items, resulted in an overestimation of item discrimination, while a negative interaction resulted in an underestimation of item discrimination. This effect was moderated by (a) the difficulty of the interacting item: more extreme difficulty values led to less bias; and (b) the number of items: the more items, the smaller the bias. In addition, violation of local independence increased the rejection rate of the likelihood ratio test. Effects were larger for $n = 1000$ compared to $n = 500$, but result patterns were similar.

### 3.3.6 Nonparametric IRT

Confirmatory application of NIRT has not been subject to much simulation research. Work of Hemker and Sijtsma (1995), Mroch and Bolt (2006), and Van Abswoude, Van der Ark, and Sijtsma (2004) was focused on dimensionality assessment in exploratory scale analysis. De Gruijter (1994) illustrated that parametric and nonparametric IRT can be compared within the framework of latent class analysis, focusing on the IRFs. Van Onna (2004) investigated estimation of the sampling distribution of Loevinger's $H$ under various conditions. Monte Carlo simulation has not been used, so far, to evaluate confirmatory application of NIRT, let alone to compare NIRT with parametric IRT.

## 3.4 Discussion

In the following, the results of the literature discussed are translated to a set of expectations and the corresponding design of our Monte Carlo simulation study, which is described further in the subsequent chapters. In Tables 3.1a and 3.1b an overview is presented of the main design and performance variables of the reviewed previous research in chronological order of publishing.

*Table 3.1a.* Design characteristics of simulation studies, presented in chronological order of publication.

| Study | Abbr.[a] | R[b] | n[c] | Item type[d] | #Items | Item dist.[e] | $\lambda$ ($\alpha$)[f] | #LVs | LV dist.[g] |
|---|---|---|---|---|---|---|---|---|---|
| B. O. Muthén & Kaplan, 1985 | MK85 | 25 | 1000 | P | 4 | n/ms/s/ hs/hk | 0.7 | 1 | n |
| Babakus et al., 1987 | B87 | 300 | 100–500 | P | 4 | n/s/hs/b | 0.4–0.8 | 1 | n |
| Drasgow, 1989 | D89 | 10 | 200–1000 | D | 5–25 | *lv* | (0.4)–(1.8) | 1 | n |
| Knol & Berger, 1991 | KB91 | 10 | 250–1000 | D | 15–30 | *lv* | (1.0)–(1.5) | 1–3 | n |
| Parry & McArdle, 1991 | PM91 | 31 | 50–200 | D | 8 | n/s/hs | 0.5–0.9 | 1 | n |
| B. O. Muthén & Kaplan, 1992 | MK92 | 1000 | 500–1000 | P | 4–15 | n/ms/s hs/hk/u | 0.7 | 1–3 | n |
| Stone, 1992 | S92 | 100 | 250–1000 | D | 10–40 | *lv* | (0.7)–(3.0) | 1 | n/s/nk |
| Potthast, 1993 | P93 | 100 | 500–1000 | P | 4–16 | n/s hs/u | 0.7 | 1–4 | n |
| Dolan, 1994 | D94 | 100 | 200–400 | D/P | 8 | n/s | 0.7–0.9 | 1 | n |
| Maydeu-Olivares et al., 1994 | M94 | 3 | 250–1000 | P | 5–25 | *lv* | *n.s.* | 1 | n |
| Boulet, 1996 | B96 | 100 | 250–10000 | D | 15–60 | *lv* | (0.5)–(1.5) | 1 | n/s/hs |
| Coenders et al., 1997 | C97 | 200 | 1000 | D/P | 4 | *lv* | 0.9 | 2 | n/ms/s |
| Parshall et al., 1997 | P97 | 100 | 100–1000 | D | 80 | *lv* | *n.s.* | 1 | n |
| Finger, 2001 | F01 | 5 | 250–2000 | D | 10–50 | *lv* | (0.7)–(1.6) | 1 | n |
| Tuerlinckx & De Boeck, 2001 | TD01 | 50 | 500–1000 | D | 6–20 | *lv* | (1.0) | 1 | n |
| DiStefano, 2002 | D02 | 100 | 350–700 | P | 12–16 | n/hs | 0.3–0.7 | 2–3 | n |
| Tate, 2003 | T03 | 1 | 2000 | D | 60 | *lv* | 0.5–0.8 | 1–4 | l |
| Flora & Curran, 2004 | FC04 | 500 | 100–1000 | D/P | 5–20 | *lv* | 0.7 | 1–2 | n/ms mk/s/k |
| Kay, 2004 | K04 | 5 | 50–1000 | D | 30–100 | *lv* | (0.7) | 1 | n/u |
| Beauducel & Herzberg, 2006 | BH06 | 500 | 250–1000 | D/P | 5–40 | *lv* | 0.5–0.6 | 1–8 | n |
| Forero & Maydeu-Olivares, 2009; Forero et al., 2009 | F09 | 1000 | 200–2000 | D/P | 9–42 | n/s | 0.4–0.8 | 1–3 | n |
| Trierweiler, 2009 | T09 | 500 | 100–1000 | P | 9–18 | n/ms/s | 0.6 | 3 | n |
| DeMars, 2010 | D10 | 1000 | 300–5000 | D | 45 | *lv* | (0.3)–(1.3) | 1 | s/k |
| Finch, 2010/2011 | F10 | 1000 /500 | 250–2000 | D | 15–60 | *lv* | (0.4)–(2.0) | 2 | n/s |
| Holgado-Tello et al., 2010 | H10 | 100 | 1000 | P | 12 | n/s | 0.4–0.7 | 3–5 | n |
| Yang-Wallentin et al., 2010 | Y10 | 2000 | 100–1600 | D/P | 6–16 | n/s | 0.6–0.9 | 2–4 | n |
| Maydeu-Olivares et al., 2011 | M11 | 1000 | 500–5000 | D/P | 10–20 | n | 0.4–0.8 | 1 | n |
| Rhemtulla et al., 2012 | R12 | 1000 | 100–600 | D/P | 10–20 | n/ms/s/hs | 0.3–0.7 | 2 | n/s |

[a] Abbreviation used to refer to this study in subsequent tables. [b] Number of replications $R$. [c] Minimum and maximum sample size included. [d] D: dichotomous, P: polytomous item variables. [e] Item distributions included; n: approximately normal, ms: mildly skewed, s: moderately skewed, hs: highly skewed, mk: mildly kurtotic, k: moderately kurtotic, hk: highly kurtotic, nk: negative kurtosis, pk: positive kurtosis, b: bimodal, u: uniform, l: logistic, *lv*: item distribution is not manipulated independently of the latent variable. [f] Minimum and maximum loading (discrimination) parameter value included; *n.s.*: not specified. [g] Latent variable distributions included, see (e) for explanation of abbreviations.

*Table 3.1b.* Additional design characteristics and performance variables of simulation studies, presented in chronological order of publication.

| Study | Estimation models | Software[a] | Performance variables | | | |
|---|---|---|---|---|---|---|
| | | | Parameters | Standard errors | Model fit | ANOVA |
| MK85 | FA-lin-ML, FA-lin-GLS, FA-lin-WLS/ADF, FA-tet-CVM | LISCOMP | bias | bias | $\chi^2$ | |
| B87 | FA-lin-ML, FA-poly-ML, FA-rho-ML, FA-tau-ML | LISREL | bias, MSE | bias | $\chi^2$, GFI, AGFI, RMR | |
| D89 | IRT-2p-MML, IRT-2p-JML | $pc\times 2$[b] | bias | bias | | |
| KB91 | FA-tet-IP, FA-tet-ULS, FA-tet-GLS, FA-tet-ML, FA-tet-$\alpha$, FA-tet-MINR$_{\text{adj}}$, FA-pa-ULS, IRT-2p-MML, IRT-2p-JML | SPSS$^{\text{x}}\times 5$, $pc$, NOHARM, TESTFACT, MAXLOG | MSE | | | |
| PM91 | FA-phi-ULS, FA-tet-ULS, FA-tet-WLS, FA-pa-ULS | SAS$\times 2$, LISCOMP, NOHARM | bias | | | |
| MK92 | FA-lin-GLS, FA-lin-WLS/ADF | LISCOMP | bias | bias | $\chi^2$ | |
| S92 | IRT-2p-MML | MULTILOG | bias, RMSE | | | |
| P93 | FA-poly-CVM | LISCOMP | bias | bias | $\chi^2$ | |
| D94 | FA-lin-ML, FA-poly-ML, FA-poly-WLS, FA-poly-GLS | PRELIS/LISREL$\times 3$, LISCOMP | bias | bias | $\chi^2$, AGFI | |
| M94 | IRT-grm-ML, IRT-pc2-ML | MULTILOG | | | IOI | $\surd$ |
| B96 | IRT-2p-MML, FA-pa-ULS | TESTFACT, NOHARM | bias, RMSE | bias | | $\surd$ |
| C97 | FA-lin-ML, FA-poly-WLS, FA-poly-CVM | LISREL, PRELIS/LISREL$\times 2$ | bias | | | |
| P97 | IRT-1p-MML, IRT-2p-MML, IRT-3p-MML | BILOG | MSE, RMSE, $\rho(\hat{\omega},\omega)$[c], SD | bias | CV | |
| F01 | IRT-2p-MML, FA-pa-ULS, FA-tet-ULS, FA-phi-ULS | $pc\times 4$ | bias, MSE, RMSE, $r(\hat{\omega},\omega)$[d] | | | $\surd$ |
| TD01 | IRT-2p-MML | MULTILOG | bias | bias | LRF | |
| D02 | FA-lin-ML, FA-poly-WLS | PRELIS/LISREL | bias | bias | $\chi^2$, GFI, NNFI, SRMR, RMSEA | |
| T03 | EFA-tet-ULS, FA-tet-WLSMV, FA-pa-ULS, IRT-2p/3p-MML, IRT-2p/3p-LID, IRT-np-HCA, IRT-np-DIM, IRT-np-DET | MPLUS$\times 2$, NOHARM, CHIDIM, TESTFACT, IRT-LD, HCA, DIMTEST, DETECT | estimate | | | |
| FC04 | FA-poly-WLS, FA-poly-WLSMV | MPLUS | bias | bias | $\chi^2$ | |
| K04 | FA-tet-ML, IRT-2p-MML | PRELIS/LISREL, BILOG | bias, MSE, RMSE, $r(\hat{\omega},\omega)$ | | | $\surd$ |
| BH06 | FA-lin-ML, FA-poly-WLSMV | MPLUS | bias | mean | $\chi^2$, CFI, TLI, RMSEA, SRMR | $\surd$ |
| F09 | IRT-grm-ML, FA-poly-ULS, FA-poly-WLSMV | MPLUS | bias, RMSE | bias | | |
| T09 | FA-lin-ML, FA-poly-WLS, FA-poly-DWLS, FA-poly-MLR | PRELIS/LISREL | bias, RMSE | bias | $\chi^2$, RMSEA, CFI | |
| D10 | FA-pa-ULS, FA-tet-WLSMV, IRT-2p-MLR | NOHARM, MPLUS$\times 2$ | bias, SD | bias | | |
| F10 | FA-pa-ULS, FA-tet-WLSMV, IRT-2p-MML | NOHARM, MPLUS, BILOGMG | bias, RMSE | mean | | $\surd$ |
| H10 | FA-lin-ML, FA-poly-WLS | PRELIS/LISREL | | | $\chi^2$, GFI, AGFI, RMSEA | |
| Y10 | FA-poly-ML, FA-poly-ULS, FA-poly-WLS, FA-poly-DWLS | PRELIS/LISREL | bias, RMSE | bias, RMSE | $\chi^2$ | |
| M11 | FA-lin-ML, IRT-grm-ML | $pc$ | | | $\chi^2_B$, $\chi^2_{YB}$, $F_{YB}$ | |
| R12 | FA-lin-ML[e], FA-poly-WLSMV | MPLUS | bias | bias | $\chi^2$ | |

[a] $pc$: Model was estimated using personal code. [b] $\times x$ denotes that the software was used to estimate $x$ models.
[c] $\rho$ refers to Spearman's rank correlation coefficient. [d] $r$ refers to Pearson's product-moment correlation coefficient.
[e] Accompanied by robust corrections for nonnormality.

### 3.4.1 Setup of the Monte Carlo Study

The main explanatory factors of the design are presented first, followed by the elements that are held constant in the study.

#### Estimation Model

Four models are included in our comparisons:

1. FA of the sample covariance matrix with maximum likelihood estimation (e.g., Jöreskog, 1967; *FA-lin*);
2. FA of the estimated polychoric correlation matrix (Olsson, 1979) using mean-and-variance adjusted WLS (B. O. Muthén, 1984; *FA-poly*);
3. The graded response IRT model (Samejima, 1969) with robust ML (L. K. Muthén & Muthén, 1998–2010, p. 533; *IRT-grm*);
4. The nonparametric Mokken IRT model (Mokken, 1971) extended to polytomous items (I. W. Molenaar, 1982; *IRT-mok*).

All models are applied in a confirmative sense, i.e., to test specific hypotheses about the structure of the data. FA-lin-ML is included as the standard practice. Although, by definition, the FA-lin model does not hold for discrete item variables, it is still applied oftentimes Chapter 2. It is therefore of interest to investigate the robustness of the model against commonly found distributional anomalies.

The reason for including both FA-poly and IRT-grm, even though they have been shown to be theoretically equivalent (Takane & De Leeuw, 1987) is twofold. First, the models are typically estimated using different estimation methods. Second, the theoretical equivalence holds under the assumption of normality. As we are interested in nonnormal item and/or LV distributions, conditions in which the theoretical equivalence does not necessarily hold, both models are included in our comparisons. Furthermore, both WLSMV and MLR seem rather promising in their robustness to violations of normality (e.g., Forero & Maydeu-Olivares, 2009; Forero et al., 2009; Trierweiler, 2009; Yang-Wallentin et al., 2010).

IRT-mok is included, because its assumptions are weaker than those of the parametric models, making it — presumably — more suitable for nonnormal data than the other models. Furthermore, nonparametric and parametric IRT or FA models have not been compared thoroughly for scale analysis. As preliminary results indicate that nonparametric IRT could be a good alternative to parametric models in case of nonnormal item distributions (Tate, 2003), investigating the generalizability of these results is of great importance. We chose to include the Mokken model extended to polytomous items, because its main result, the scaling coefficient Loevinger's $H$, provides a useful means of crossing the bridge between parametric and nonparametric models: It gives an indication of the strength of an item in representing the LV, similar to factor loadings and discrimination parameters.

### Latent Variable and Item Response Distribution

Ability and achievement score distributions are often multimodal or skewed, rather than normal, as was found by Micceri (1989), who investigated the distributional properties of empirical samples in the social sciences. Hence, the assumption of a normally distributed LV is often untenable. In addition, categorical items can at best *approximate* a normal distribution, because of their discrete nature. Our principal interest is to investigate the effect of nonnormal distributions of LVs, item variables, or both, on model estimation.

The skewness of ordered categorical variables can be the result of (a) a skewed LV, via skewed underlying continuous item variables, or (b) the location of the threshold parameters. Results thus far seem to indicate that for FA-poly-WLSMV/ DWLS the former source of nonnormality has a stronger adverse effect on model estimation than the latter (Flora & Curran, 2004; Forero et al., 2009; Rhemtulla et al., 2012; Trierweiler, 2009; Yang-Wallentin et al., 2010). Furthermore, in case of nonnormal LV distributions, FA produced inferior results compared to IRT (Boulet, 1996; DeMars, 2010; Finch, 2010; Kay, 2004). Whereas, in case of nonnormal item distributions (combined with normal LVs), FA-poly as estimated by CVM, WLSMV/DWLS, or ULS performed rather well for skewness values up to about 1.5 (DiStefano, 2002; Forero et al., 2009; Potthast, 1993; Rhemtulla et al., 2012; Trierweiler, 2009; Yang-Wallentin et al., 2010). To our knowledge, IRT-grm performance has not been investigated yet. For IRT-2p-MML, however, Stone (1992) found parameters to be overestimated when the LV distribution was skewed or platykurtic.

Both Coenders et al. (1997) and Rhemtulla et al. (2012) found that, when thresholds were nonequally spaced, violating the FA-lin-ML assumption of a linear relationship between the LV and the observed variables, FA-lin-ML estimators of loadings were most biased, regardless of the LV distribution. As a result, FA-lin performed worse than FA-poly in case of a normal LV and nonnormal items, but better when the LV was skewed as long as thresholds were spaced evenly.

A systematic study of the effect of both types of skewness would contribute to a full understanding of the consequences of LV and item skewness on model estimation. We therefore include a normal and a skew-normal LV distribution in our design, as well as both normal and skewed item distributions. A condition of a scale consisting of both left- and right-skewed items is also included, because the combination of left- and right-skewness is expected to worsen polychoric correlation estimation as a result of an increased probability of small cell frequencies in item cross tables. In addition, a bimodal item distribution is included, representing, for example, the rather common occurence of items respondents are inclined to either strongly agree or strongly disagree with.

### Strength of the Item Response Scale

Scales ideally consist of items that are all strongly related to the LV of interest, because this facilitates LV score estimation. However, in practice, items comprising a scale

are often heterogeneous in the strength of their relationship with the LV. Therefore, we include a condition of all high loadings and a condition of mixed-loadings in our design.

### Sample Size

In general, a larger sample size results in decreased estimation variability. For large sample sizes ($n \geq 1000$), model estimation is rather robust to the violation of distributional assumptions. As we are interested in model performance for sample sizes representing the lower end of what is found in practice (see Chapter 2), we include sample sizes of 200 and 600.

### Number of Item Response Categories

In many simulation studies only dichotomous items were studied. Hence, the generalizability of those results to polytomous items remains to be investigated. Our study is concerned with polytomous items, and restricted to five-category item variables because this number of categories is common for items of scales in the social and behavioral sciences, as was found in Chapter 2. Moreover, it was found (e.g., Dolan, 1994) that the standard practice of applying a linear factor model requires at least five response categories.

### Number of Items and Latent Variables

No effect was found of test length on performance variables, when the number of items was approximately 10 or more and the LV structure was correctly specified in the model (e.g., Boulet, 1996; Drasgow, 1989; Finch, 2010, 2011; Finger, 2001; Kay, 2004; Knol & Berger, 1991). When the item/LV ratio was as small as three — or at least smaller than seven — model estimation deteriorated, and more so for FA than for IRT models (Forero & Maydeu-Olivares, 2009; Forero et al., 2009). Stone (1992), however, found a decrease in parameter bias and RMSE in case of a skewed or platykurtic LV distribution as the number of items increased from 10 to 40 for IRT-2p-MML.

To avoid estimation problems related to test length, the number of item variables is held constant at 12 in our study, exceeding the number most commonly found in practice (see Chapter 2). We chose a unidimensional population model for reasons of simplicity and because in practice many scales are unidimensional by theory and construction. Consequently, the resulting item/LV ratio is also sufficiently high.

### Design Summary

In summary, the design has the following explanatory factors.

- Varying design factors:
  - Estimation model (FA-lin-ML, FA-poly-WLSMV, IRT-grm-MLR, IRT-mok)

  - LV distribution (normal, left skew-normal)
  - Scale shape (various combinations of normal, right-skewed, left-skewed, and bimodal item response distributions)
  - Scale strength (strong: all items load strongly [0.80] on the LV; or mixed: 4 items strong [0.80], 4 medium [0.50], and 4 weak [0.30])
  - Sample size (small: $n = 200$; medium: $n = 600$)

- Constant design factors:

  - Number of item response categories (5)
  - Number of items (12)
  - Number of LVs (1)

In evaluating the scaling models' performance, we are particularly interested in the properties of estimators of item parameters and corresponding standard errors, model fit, and LV score distributions. The response or performance variables, as well as their evaluation criteria, are elaborated in Chapter 4.

LV score estimation has typically been ignored in FA simulation studies, presumably because the estimation of factor scores is often omitted in the practice of FA, and researchers applying FA heuristically use simple sum scores to estimate their respondents' LV scores. However, as LV score estimation is common practice in IRT modeling, and obtaining LV scores is, generally, often the main objective of scale development, we include the evaluation of LV estimators in our comparison of the estimation models.

### 3.4.2 Expectations

Our main expectations regarding the anticipated results from the presented Monte Carlo design based on the literature are sketched very briefly here; a detailed specification of our hypotheses with respect to the performance variables is given in Chapter 4.

Under conditions of normality, we do not expect much difference in the performance of the estimation models. Loading parameters are expected to be estimated quite accurately by FA-poly and IRT-grm (e.g., Forero & Maydeu-Olivares, 2009) and underestimated by FA-lin (e.g., Trierweiler, 2009). Standard errors are thought to be negatively biased for the small sample size for FA-lin (Rhemtulla et al., 2012) and unbiased for FA-poly (e.g., Flora & Curran, 2004) and IRT-grm (Forero & Maydeu-Olivares, 2009). We expect that model fit indices behave properly for all models. Estimated LV scores might be somewhat attenuated towards the mean of the distribution, especially for FA-lin and simple sum scores (Dumenci & Achenbach, 2008), the latter of which are taken as the IRT-mok LV score estimates in our study.

When LV and/or item distributions deviate from normality, we expect to observe larger differences in model performance. Generally, IRT-grm is expected to perform

well in terms of parameter estimation in harsh conditions such as nonnormality and a small sample size (Forero & Maydeu-Olivares, 2009). FA-poly parameter estimation will be more adversely affected by nonnormality of the LV distribution than by nonnormality of item distributions (e.g., DeMars, 2010; DiStefano, 2002). FA-lin parameter estimators are expected to be severely biased under nonnormal conditions, except when item thresholds are evenly spaced (e.g., Coenders et al., 1997). We believe that IRT-mok might provide a useful alternative to the parametric models in these conditions. Standard error estimators are expected to be biased for FA-lin (e.g., DiStefano, 2002) and FA-poly (e.g., Flora & Curran, 2004), but not for IRT-grm (DeMars, 2010). As the model fit indices RMSEA and SRMR have been found to behave properly under conditions of nonnormality (e.g., Trierweiler, 2009), we expect to observe this robustness for the included parametric models. LV score estimation is thought to suffer from nonnormal LV and item distributions (Dumenci & Achenbach, 2008), although it has not been thoroughly investigated in previous research.

In addition, we expect more accurate LV scores and more appropriate fit measures in the strong-scale condition than in the mixed-scale condition (e.g., Forero & Maydeu-Olivares, 2009) for all models included in our design. Finally, we anticipate better results for the medium than for the small sample size.

In the next chapter, the setup of our Monte Carlo simulation design is presented. It includes a detailed description of the procedures used for data generation and the criteria applied for the evaluation of the results by means of performance variables. Finally, a detailed set of hypotheses is given based on the literature that was discussed in the present chapter.

# Chapter 4

# Setup of the Simulation Study

Much research has already been conducted on the robustness of factor analysis (FA) and item response theory (IRT) models. In our Monte Carlo study we aim to replicate some of those earlier findings, bringing together the results from the separate fields of FA and IRT, and to contribute by systematically comparing the models' robustness against combinations of latent variable (LV) and item skewness. We use four methods of scale analysis under various distributional conditions: FA of the sample covariance matrix (FA-lin), FA of the estimated polychoric correlation matrix (FA-poly), the graded response IRT model (IRT-grm), and the nonparametric Mokken IRT model (IRT-mok). By taking into account a full spectrum of performance variables, i.e., estimators of parameters and corresponding standard errors, model fit indices, and LV scores, we provide a broad evaluation of the models' robustness against violations of assumptions under investigation. In contrast to the abundant literature on, mainly, FA models, IRT-mok has not been subject to much simulation research. Thus, by including the nonparametric IRT-mok model in our comparative study, we aim to shed some light on the performance of IRT-mok under conditions of LV and item skewness itself, and in comparison with the parametric models.

In the current chapter the setup of the simulation study is described. The data generation process is explained, and the performance variables and applied criteria for evaluating the results are addressed. Finally, our expectations are presented with regard to the performance variables of the simulation study, leaning on the literature discussed in the previous chapter.

In the subsequent two chapters the results of the study are presented and discussed. Chapter 5 comprises the results regarding normal data configurations, thus generating a frame of reference for the results presented in Chapter 6, regarding the data configurations that represent the more serious and realistic model violations we are interested in,

# 4.1 Data Generation Proces

Several steps are taken in the data generation process. The population model, according to which all data are generated, is explained first. We then focus on the various parameterizations common to FA and IRT, and recapitulate the relations between FA and IRT parameters. The method of standardizing parameters and standard errors is clarified next. We conclude this section with a detailed description of the data generation steps and a justification of the chosen number of replications.

## 4.1.1 Population Model

Recall from Chapter 1 our general definition of the polychoric factor model,

$$\mathbf{X}^* = \boldsymbol{\theta}\boldsymbol{\lambda}' + \mathbf{E}, \tag{4.1}$$
$$X_{is} = c \quad \text{for} \quad \tau_{ic} < X^*_{is} < \tau_{i(c+1)},$$

with $c = 0, 1, \ldots, C$, $\tau_{i0} = -\infty$, $\tau_{iC} = \infty$, $i = 1, 2, \ldots, I$, and $s = 1, 2, \ldots, n$. $\mathbf{X}^*$ $(n \times I)$ is a matrix of *latent* continuous item scores for respondents $s$ on items $i$, $\boldsymbol{\lambda}$ is a vector of item loadings of length $I$ on the latent dimension, $\boldsymbol{\theta}$ is the vector of LV scores of length $n$, $\mathbf{E}$ $(n \times I)$ is an error matrix with independent normally distributed elements $\epsilon_{is}$, $X_{is}$ are the *observed* categorical item scores, $\tau_{ic}$ are the thresholds of item $i$ and category $c$, and $C$ is the number of response categories. Data are generated according to this model.

### LV Distribution

Under the polychoric factor model, the LV has a normal distribution. Because we are interested in model performance in case of a skew-normal LV as compared to the condition of a normal LV, data are also generated using a skew-normal LV.

We chose a distribution from Azzalini's (1985) class of skew-normal distributions to represent the skew-normal LV, because it allows for an optimal comparison with the normal distribution. In his Equation 3, Azzalini defines the skew-normal density function for a skew-normal random variable $z$ with parameter $\iota$, or $z \sim \mathcal{SN}(\iota)$ as

$$\phi(z; \iota) = 2\phi(z)\,\Phi(\iota z), \qquad -\infty < z < \infty, \tag{4.2}$$

where $\phi$ and $\Phi$ denote the standard normal densitiy and distribution function, respectively, and $\iota$ is the shape parameter. The standard normal distribution is retrieved as $\mathcal{SN}(0)$. Azzalini (2005) extends this class of distributions to the multivariate case, additionally specifying a scale and location parameter, which affect the variance and location of the distribution, respectively.

In Figure 4.1 both LV distributions included in our study are shown: $\mathcal{SN}(-2.61, 3.29, 10)$ and $\mathcal{N}(0, 4)$. The parameters of the skew-normal distribution are *location*, *scale*, and *shape*, respectively. The shape parameter determines the skewness of the distribution and its sign is congruent with the sign of the skewness;

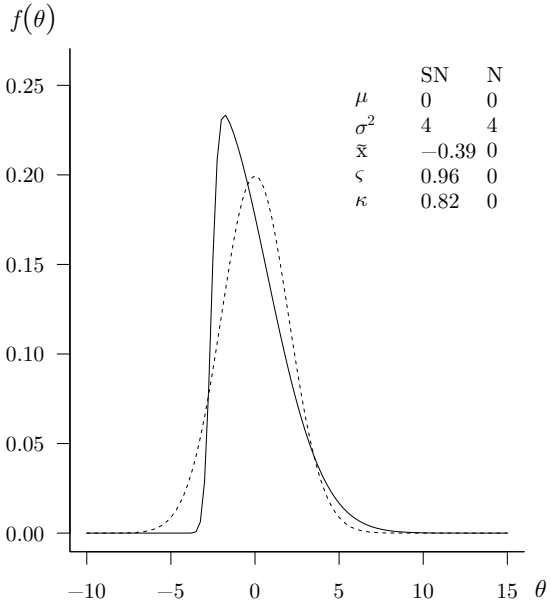



*Figure 4.1.* Distribution of a right skew-normal $\mathcal{SN}(-2.61, 3.29, 10)$ (solid line) and a normal $\mathcal{N}(0, 4)$ (dashed line) latent variable. The mean, variance, median, skewness, and excess kurtosis of $\theta$ in the respective distributions are shown in the inset.

we set it to a value of 10 to result in a moderately skewed distribution with a skewness of about 0.96. The location and scale parameter are set such that, given the shape parameter, the mean and variance match those of the included normal distribution $\mathcal{N}(0, 4)$.

### Item Distribution

The distribution of ordered categorical variables is a result of the underlying LV distribution and the location of the threshold parameters. As we aim to systematically investigate the effect of the LV and item distributions, we manipulate the threshold values to accomplish the various combinations of LV and item skewness.

The thresholds $\tau_{ic}$ determine the values of $\mathbf{X}$ $(n \times I)$, the matrix of observed ordered categorical item scores, and are set to create four shapes of item distributions for $\mathbf{X}$: normal, bimodal, right-skewed, and left-skewed, as is illustrated in Figure 4.2. The distribution of the normal items is chosen to result in a zero skewness and excess kurtosis. The bimodal item category proportions are set to result in a symmetric distribution with two modes. The skewed items are composed to have one clear asymmetric mode, and to result in an absolute skewness of about 1.5 which has been

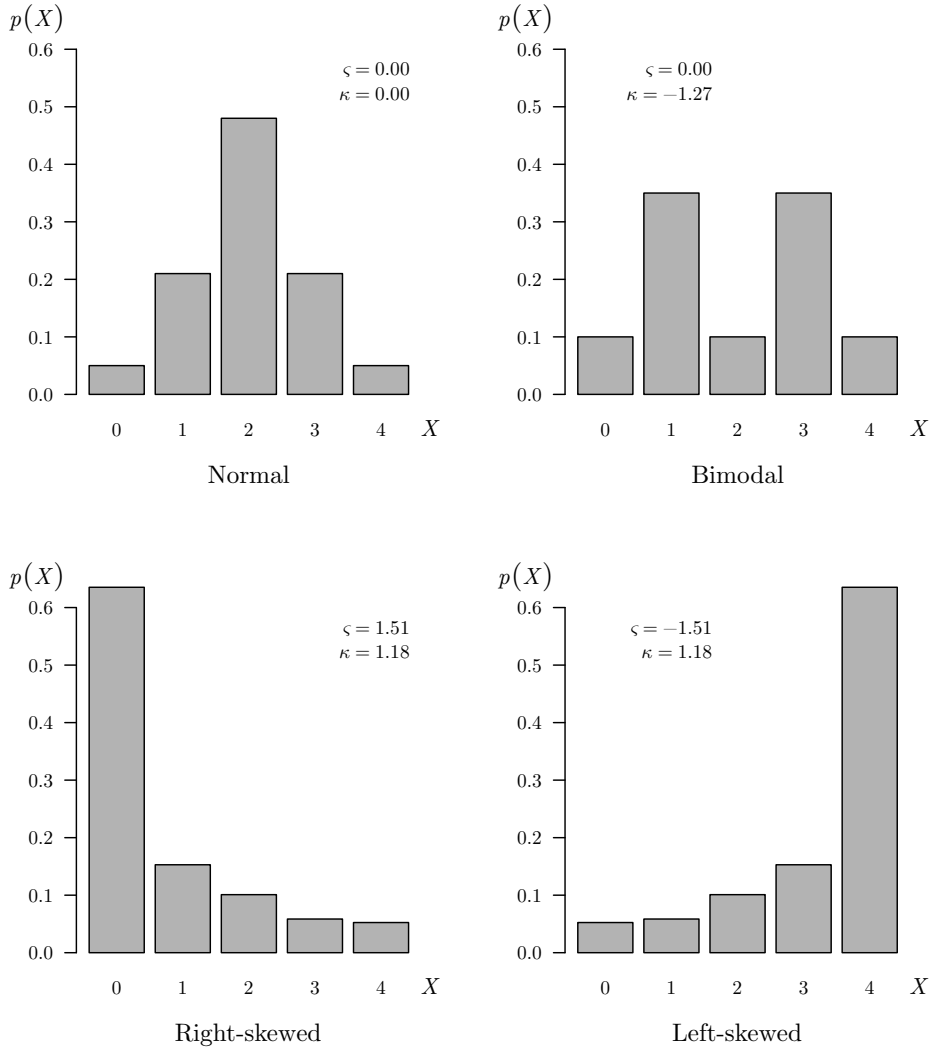



*Figure 4.2.* Four types of item distributions. The skewness and excess kurtosis of $X$ in each distribution type are given in the inset.

considered a border point between moderate and severe skewness in previous studies (e.g., Forero & Maydeu-Olivares, 2009).

The values of the thresholds to be set result in a specific proportion of the respondents being in each category of the discrete item. Therefore, the thresholds are quantiles of the distribution of the latent continuous item, corresponding to these various proportions. For items loading on the normal LV, $\tau_{ic}$ is the normal quantile that corresponds to the area under the standard normal curve, of a size that equals the cumulative proportion of all categories up to $c$. For items loading on the skew-normal LV, thresholds are determined similarly, using a skew-normal distribution[a]. By taking this approach, two identically distributed items can be created, though one loads on a normal LV and the other on a skew-normal LV, as a result of their different threshold values. All threshold values used in our simulation study are listed in Appendix C.1.

In Appendix C.2 an illustration of the relations between LV scores, latent item scores, observed item scores, and scale scores is provided, demonstrating the effect of the threshold values on the transformation of $X^*_{is}$ into $X_{is}$. In the example we show how the sum of a number of normal ordered categorical item variables can have a skew-normal distribution as a result of a skew-normal underlying LV combined with a specific set of thresholds.

Given the polychoric factor model (Equation 1.7), the population covariance matrix $\mathbf{\Sigma}^*$ $(I \times I)$ for the latent item scores $\mathbf{X}^*$ is defined as

$$\mathbf{\Sigma}^* = \boldsymbol{\lambda}\phi\boldsymbol{\lambda}' + \mathbf{\Psi}, \tag{4.3}$$

with $\boldsymbol{\lambda}$ $(I \times 1)$ the vector of loadings, $\phi$ the LV variance, and $\mathbf{\Psi}$ $(I \times I)$ the error covariance matrix.

The derivation of the model-implied covariance matrix for the observed item scores $\mathbf{X}$ is more complex. We followed Maydeu-Olivares et al. (2011), who presented the formulas for obtaining the variance and covariance for IRT-grm, given the model parameters. For FA-poly, we used that same reasoning, applying the formulas to the IRT formulation of FA-poly taken from Takane and De Leeuw (1987). The formulas are provided in Appendix C.3. All integrals are approximated by summing over discrete values of the LV distribution.

### 4.1.2 Parameterization

Although data are generated under the FA-poly model, we could equivalently have used an IRT parameterization. In the following, the IRT-grm parameters and the IRT-mok coefficients are discussed.

[a] Latent continuous item variables loading on a skew-normal LV with a normal error term follow a distribution that is a convolution of a skew-normal and a normal distribution, which is in itself a skew-normal distribution. The location, shape, and scale parameter of this distribution are functions of the parameters of the originating skew-normal and normal distributions (Azzalini, 1985; Henze, 1986).

### IRT-grm Parameters

FA and IRT each have their own set of parameters: loadings $\lambda$ and thresholds $\tau$ for FA, and discrimination $\alpha$ and (step-)difficulty $\beta$ for IRT. The shared parameters are LV variance $\phi$ and error variance $\psi$. IRT parameters can be computed from FA parameters and vice versa, using the well-known conversion formulas (e.g., Takane & De Leeuw, 1987),

$$\alpha_i^{\mathcal{N}} = \frac{\lambda_i}{\sqrt{\psi_i}}, \tag{4.4}$$

and

$$\beta_{ic} = \frac{\tau_{ic}}{\lambda_i}, \tag{4.5}$$

where $\alpha_i^{\mathcal{N}}$ is the discrimination parameter for item $i$ in the normal-ogive model, $\psi_i$ the error variance, and $\beta_{ic}$ the step-difficulty parameter of step $c$ for item $i$. Equations 4.4 and 4.5 hold for the normal-ogive form of the IRT-grm model, see Equation 1.20. For a correct conversion of parameters $\lambda_i$ and $\psi_i$ into the discrimination parameter $\alpha_i$ of the logistic-ogive form of the IRT-grm model, see Equation 1.21, an adjustment to Equation 4.4 is required. Since the normal-ogive and logistic-ogive function are nearly equivalent when a scaling factor $d \equiv 1.702$ is used (see, e.g., Camilli, 1994; I. W. Molenaar, 1974), we apply that factor to accommodate the logit scale of the logistic IRT-grm:

$$\alpha_i = d\frac{\lambda_i}{\sqrt{\psi_i}}. \tag{4.6}$$

### IRT-mok Coefficients

The fourth model under investigation, IRT-mok, is the monotone homogeneity model extended to polytomous items (Mokken, 1971; I. W. Molenaar, 1991). Recall Equation 1.36, for scalability coefficient Loevinger's $H$:

$$H_{ij} = 1 - \frac{F_{ij}}{E_{ij}},$$

with $F_{ij}$ the total weighted number of Guttman errors for item pair $(i, j)$ and $E_{ij}$ the *expected* total weighted number of Guttman errors if the items were independent. Also recall that the item coefficient $H_i$ and scale coefficient $H_{scale}$ can be derived from $F_{ij}$ and $E_{ij}$; see Equations 1.37 and 1.38.

For the present simulation study, population values of the scalability coefficient are obtained as a reference for the evaluation of sample $H$ values. As both $F_{ij}$ and $E_{ij}$ can be computed based on the bivariate probability table of items $i$ and $j$, population values of Loevinger's $H_{ij}$, $H_i$, and $H_{scale}$ can be computed given the population bivariate probability tables of item pairs. These tables can be computed based on $\lambda_i$ and $\tau_{ic}$. To this end, the normal-ogive IRT model (Equation 1.20) is applied to the IRT transformations (Equations 4.4 and 4.5) of the population parameters $\lambda_i$ and

$\tau_{ic}$. We approximate the bivariate probability of responding $c$ to item $i$ and $d$ to item $j$ by taking the sum over the vector of all possible values of $\boldsymbol{\theta}$ weighted by the corresponding density of each particular $\theta_s$,

$$P(X_i = c \wedge X_j = d) \approx \sum_s w_{\theta_s} \frac{1}{\sqrt{2\pi}} \int_{\alpha_i^{\mathcal{N}}(\theta_s - \beta_{i(c+1)})}^{\alpha_i^{\mathcal{N}}(\theta_s - \beta_{ic})} e^{-t^2/2} dt \, \frac{1}{\sqrt{2\pi}} \int_{\alpha_j^{\mathcal{N}}(\theta_s - \beta_{j(d+1)})}^{\alpha_j^{\mathcal{N}}(\theta_s - \beta_{jd})} e^{-t^2/2} dt, \quad (4.7)$$

with weights $w_{\theta_s}$ taken from the normal or the skew-normal density function, depending on the applied LV distribution.

### 4.1.3 Standardization

Parameters can be put in standardized form to be independent of the — often arbitrary — metrics of the LV, the latent item variables, and the observed item variables. As a result, standardized parameters can be compared more easily across studies. FA and IRT each have a typical standardization (cf. Mehta & Taylor, 2006). In FA, parameters are often standardized to unit LV variance $\phi$ and unit latent item variance $\sigma_{i*}^2$, but other standardizations are also employed. In IRT, error variance $\psi_i$ is always set to 1; in addition, parameters are often standardized to unit LV variance $\phi$.

The conversion formulas in Equations 1.43 and 4.5 are only applicable to the unstandardized parameters. When standardized parameters are needed, the FA and IRT parameters are standardized according to their respective conventions. As we focus on standardized parameters, we denote the *standardized* parameter simply by $\omega$, whereas the *unstandardized* parameter is denoted by $\omega^\circ$.

Loadings are standardized by multiplying by the LV standard deviation $\sqrt{\phi^\circ}$ and dividing by the latent item standard deviation $\sigma_{i*}^\circ$ (cf. T. A. Brown, 2006, p. 136),

$$\lambda_i = \lambda_i^\circ \frac{\sqrt{\phi^\circ}}{\sigma_{i*}^\circ}. \quad (4.8)$$

Thresholds are standardized simply by dividing by the latent item standard deviation,

$$\tau_{ic} = \frac{\tau_{ic}^\circ}{\sigma_{i*}^\circ}. \quad (4.9)$$

As stated above, IRT parameters are typically standardized with regard to the LV only, i.e., by setting the LV variance $\phi$ to 1, while also keeping the error variances $\psi_i$ fixed at 1,

$$\alpha_i = \alpha_i^\circ \sqrt{\phi^\circ}, \quad (4.10)$$

$$\beta_{ic} = \frac{\beta_{ic}^\circ}{\sqrt{\phi^\circ}}. \quad (4.11)$$

### 4.1.4 Data Generation Steps

The vector of LV scores is generated by drawing from one of the two distributions $\mathcal{N}(0, 4)$ or $\mathcal{SN}(2.61, 3.29, 10)$; see Figure 4.1. The loading and threshold input parameters are specified in standardized form. Standardized loading parameters are convenient, because they are independent of the arbitrary scale of the item variable, thus facilitating the interpretation of the item-LV relationship. We quantify *strong*, *medium*, and *weak* item-LV relationships by specifying standardized loading parameter values of 0.80, 0.50, and 0.30, respectively. Threshold input parameters are also in standardized form as a result of using the standardized quantile functions to generate their values.

However, the IRT parameter estimates obtained from the model estimation software MPLUS are in accordance with the usual IRT parameterization, i.e., the error variances $\psi_i$ are fixed at 1. Therefore, data have to be generated with error variances equal to 1. Consequently, it is undesirable to also fix the LV variance at 1, because the variance of the items would then be determined to a greater extent by error than by the LV with loading values smaller than 1. Items loading strongly on the LV should be determined, in majority, by that LV and not by other factors (which would be the case, when both the LV and error variance were set at 1). To have an information/noise ratio of about 4:1 as a starting point, the LV variance is fixed at 4 instead of the standardized value of 1. Each item's ultimate information/noise ratio also depends on the loading value. As a result, the data have to be generated using unstandardized parameters. Hence, the standardized loadings and thresholds given as input to the data generation procedure have to be converted to the unstandardized form before samples of data are generated.

The loadings $\lambda_i$, are converted using the following formula (cf. T. A. Brown, 2006, p. 134f.; see also Equation 4.8),

$$\lambda_i^\circ = \lambda_i \frac{\sigma_{i*}^\circ}{\sqrt{\phi^\circ}}, \tag{4.12}$$

with

$$\sigma_{i*}^\circ = \sqrt{\frac{\psi_i^\circ}{\psi_i}}, \tag{4.13}$$

and

$$\psi_i = \sigma_{i*}^2 - \lambda_i^2 \, \phi \, = 1 - \lambda_i^2, \tag{4.14}$$

with $\lambda_i^\circ$ the unstandardized loading, $\lambda_i$ the standardized loading, $\sigma_{i*}^\circ$ the standard deviation of $X_{is}^*$, $\phi^\circ$ the unstandardized LV variance (set to 4), $\psi_i^\circ$ the unstandardized error variance (set to 1), $\psi_i$ the standardized error variance, $\sigma_{i*}^2$ the standardized variance of $X_{is}^*$, and $\phi$ the standardized LV variance (the latter two are by definition equal to 1).

The thresholds $\tau_{ic}$ are converted using (cf. Equation 4.9)

$$\tau_{ic}^\circ = \tau_{ic} \sigma_{i*}^\circ. \tag{4.15}$$

In summary, first loading and threshold parameter values are specified in standardized form. These are transformed to unstandardized form, and data are generated using data generation software. Next, model parameters are estimated using estimation software. The software specifications are given in Section 4.2.1. Finally, before they are evaluated, parameter estimates are transformed to standardized form, using Equations 4.8 and 4.11.

In Appendix C.4, an illustration of the data generation steps is provided.

### 4.1.5 Number of Replications

Each data configuration is generated multiple times to approximate the sampling distributions of the statistics of interest and to reduce the influence of irrelevant sampling fluctuations in the results. The number of replications $R$ is chosen as a trade-off between accuracy and efficiency, since increasing the number of replications results in more accurate estimates while it also increases computation time.

To determine the optimal $R$, we consider the fact that we want to obtain standard errors for our parameters of interest. Strictly speaking, the use of standard errors is only justified for statistics whose distribution is approximately normal. The number of replications required for a statistic to approximate a normal distribution depends on the statistic, but when $R$ exceeds 500, this is generally the case for most statistics (Stuart & Ord, 1987, p. 327f.).

Furthermore, when standard errors are estimated for estimated parameters, we report the coverage of the 95%-confidence interval. This statistic has its own distribution with a standard deviation $\sqrt{p(1-p)/R}$, where $p$ is 0.95 (cf. Van Duijn, 1993, p. 107). Setting $R$ to 1000 results in a standard deviation of 0.0069, and thus in coverage rates accurate up to the second decimal place on average, while keeping computation time manageable. $R$ is therefore chosen to be 1000 in our study.

#### Additional Replications

When at least one of the items in a sampled data set has a response category with zero observations, the computation of the polychoric correlation between that item and the other items is not possible. A variety of solutions have been suggested in the literature. Often, empty cells are deleted listwise from the data (e.g., Olsson, 1979); another popular option is adding a dummy frequency of 0.5 to the empty cell, and adjusting the frequencies of other cells to keep the marginal frequencies intact (M. B. Brown & Benedetti, 1977). As this issue is (still) subject to debate, we chose to exclude such data sets from our analyses. Hence, they are discarded and replaced by a new replication. It is recorded and reported how often this occurs.

## 4.2 Data Analysis

Having clarified the data generation process, we now turn to the data analysis as employed in the Monte Carlo study. We start with a description of the software

used, followed by the parameters of interest in our study. Finally, the performance variables are presented for each outcome measure, and criteria by which to evaluate the results in the subsequent chapters will be set.

### 4.2.1 Software

Data are generated with our own code written in R (Version 3.0.0; R Core Team, 2013). Sampling from distributions is done using the Mersenne-Twister pseudo-random number generator (Matsumoto & Nishimura, 1998), as implemented in R. Starting values for the pseudo-random number generator are generated using the R package `random` (Eddelbuettel, 2009) that uses a true random number service based on atmospheric noise available on the internet, thus providing a nondeterministic start for the pseudo-random number generation. These values are saved for the purpose of exact replicability of our Monte Carlo study. Samples from the normal and skew-normal distribution are generated by using the standard R function `rnorm` and R package `sn` (Azzalini, 2007), respectively.

The FA-lin, FA-poly, and IRT-grm models are estimated using the MPLUS software program (Version 6.12; L. K. Muthén & Muthén, 1998–2010). The population parameter values are used as starting values for parameter estimation, as was recommended by Boomsma (1985). To ensure parameter values for FA-lin, FA-poly, and IRT-grm to be on a common scale and hence estimates to be comparable across the models, the following measures are taken: (a) using the *theta* parameterization for the FA-poly model (B. O. Muthén & Asparouhov, 2002, p. 485); (b) fixing the LV variance to a specific value for model identification; and (c) setting error variances equal to 1 in the data generation procedure. Input files for running FA-lin, FA-poly, and IRT-grm analyses on example data in MPLUS, as presented in Appendix C.4, are provided at the end of that appendix.

The nonparametric IRT-mok model is estimated in R, using the `mokken` package (Version 2.7; Van der Ark, 2011, 2007).

### 4.2.2 Parameters of Interest

Not all model parameters are of the same interest when a scale of items is analyzed. Loading parameters (in FA) and corresponding discrimination parameters (in IRT) are generally important, since they convey information on the strength of the relationship between a particular item and the LV. Item-step difficulty is typically investigated in IRT analyses, since it is indicative of the location of an item step on the LV scale: How much of the LV is needed to pass a particular item step, i.e., to be in a specific item category. Since item-step difficulty depends on both the loading and the threshold of an item, thresholds contain additional information on the source of variation in item-step difficulty.

Although the IRT-mok model does not involve any parameter estimation in the true sense, application of the model does result in a number of coefficients which can be evaluated. With the aid of these coefficients, we can attempt to compare the

*Table 4.1.* Model parameters.

| | FA-lin | FA-poly | IRT-grm |
|---|---|---|---|
| *Response variables* | | | |
| Loading $\lambda$ | est[a] (12) | est (12) | est (12) |
| Threshold $\tau$ | | est (48) | est (48) |
| Discrimination $\alpha$ | | conv[b] | conv |
| Step-difficulty $\beta$ | | conv | conv |
| *Other parameters* | | | |
| LV variance $\phi$ | fixed at 4 | fixed at 4 | fixed at 4 |
| Error variance $\psi$ | est (12) | fixed at 1 | fixed at 1 |
| Total # parameters[c] | 24 | 60 | 60 |

[a] est: estimated. [b] conv: computed using conversion formulas. [c] The number of estimated parameters for each model is indicated between brackets.

performance of IRT-mok to that of the three parametric models. We take $H_i$, as it is an indicator of the scalability of item $i$, to be closest to the loading and discrimination parameter. The item mean score over respondents $\bar{x}_i$ is taken as an indicator of item difficulty.

The parameters evaluated as outcome variables are presented in Table 4.1. Some of them are directly estimated using the aforementioned software, others are computed using formulas listed earlier.

For the evaluation of the results, standardized rather than unstandardized parameters are used. The reason for this is twofold. First, unstandardized FA-lin and FA-poly/IRT-grm parameter estimates are not comparable, as FA-lin deals with the item variables on the observed scale in the integer range $[0, 4]$, whereas FA-poly/IRT-grm unstandardized parameters are mapped on the latent scale in the continuous range $[-\infty, \infty]$. Second, the use of standardized parameter estimates facilitates the comparison with results from previous and future studies.

Results are mainly discussed in terms of FA parameters $\lambda$ and $\tau$, with the former available for all parametric models and the latter only for FA-poly and IRT-grm. In addition, to enhance comparability with previous IRT studies, IRT parameters $\alpha$ and $\beta$ are discussed to some extent for FA-poly and IRT-grm. Although $\alpha$ could be computed from $\lambda$ for FA-lin, it is not presented, because the FA-lin parameter estimates are not scaled properly unless they are standardized to unit latent item variance. Standard errors for step-difficulty parameters $\beta$ are not discussed, as they are not easily derived from the $\tau$ and $\lambda$ standard errors and do not add much to the discussion of the standard error estimates of FA parameters.

### 4.2.3 Performance Variables and Criteria

In our evaluation and comparison of the performance of the estimation models under investigation, we take a rather broad perspective, considering the quality of estimators of item *parameters*, their corresponding *standard errors*, *model fit*, and *LV scores*.

The performance variables and criteria used to evaluate each of these four aspects are described below.

**Parameters and Corresponding Standard Errors**

For the evaluation of parameter estimators, a number of performance variables are computed, comparing the estimates $\hat{\omega}$ with the true values $\omega$: the (signed) plain bias,

$$\text{PB}(\hat{\omega}) = \frac{1}{R}\sum_{r=1}^{R}\hat{\omega}_r - \omega, \tag{4.16}$$

where $R$ is the number of replications, and $\hat{\omega}_r$ is the parameter estimate for replication $r$; the (signed) relative bias,

$$\text{RB}(\hat{\omega}) = \frac{1}{R}\sum_{r=1}^{R}\frac{\hat{\omega}_r - \omega}{\omega}, \tag{4.17}$$

the root mean squared error,

$$\text{RMSE}(\hat{\omega}) = \sqrt{\frac{1}{R}\sum_{r=1}^{R}(\hat{\omega}_r - \omega)^2}, \tag{4.18}$$

which is a function of the bias of an estimator and its variance; and, finally the coverage rate of the 95%-confidence interval for parameter $\omega$, computed by taking the proportion of replications in which the estimated 95%-confidence interval contains the population value $\omega$.

In addition, the standard error estimator of a parameter estimator is evaluated, using the relative bias and the root mean squared error of the standard error estimator. Since the true value of the standard error is unknown for finite sample size, the common practice of taking the empirical standard deviation of the parameter estimates as an approximation to the true value is employed (e.g., Hoogland, 1999). Therefore, we define the RB and RMSE of the standard error estimator as follows:

$$\text{RB}(\hat{se}) = \frac{1}{R}\sum_{r=1}^{R}\frac{\hat{se}(\hat{\omega}_r) - sd(\hat{\omega})}{sd(\hat{\omega})}, \tag{4.19}$$

$$\text{RMSE}(\hat{se}) = \sqrt{\frac{1}{R}\sum_{r=1}^{R}[\hat{se}(\hat{\omega}_r) - sd(\hat{\omega})]^2}, \tag{4.20}$$

where

$$sd(\hat{\omega}) = \sqrt{\frac{1}{R-1}\sum_{r=1}^{R}(\hat{\omega}_r - \overline{\hat{\omega}})^2}, \tag{4.21}$$

and

$$\overline{\hat{\omega}} = \frac{1}{R}\sum_{r=1}^{R}\hat{\omega}_r. \tag{4.22}$$

For the evaluation of the accuracy of loading parameter estimators $\hat{\lambda}_i$, on the standardized scale of $-1$ to 1, the relative bias is examined. The reason for using the relative rather than the plain bias is that we consider a deviation of, e.g., 0.06 from a loading equal to 0.80 as less detrimental than such a deviation from a loading equal to 0.30.

For the evaluation of the accuracy of threshold parameter estimators, which are on the scale of the LV, the plain bias is used. We are interested in the plain, nonrelative, deviation of these estimators from the population value, because we consider a certain deviation from the true value important regardless of whether the threshold is in the middle of the scale or at the end.

For each performance variable, criteria are set to decide whether the parameter and standard error estimators are acceptable. The accuracy of a parameter estimator is considered acceptable when its relative bias is smaller than 0.05. Although by definition arbitrary, this five percent deviation has also been used in previous simulation studies (e.g., Hoogland, 1999).

The accuracy of a standard error estimator $\hat{se}(\hat{\omega})$ is considered acceptable when its relative bias is smaller than 0.10. This criterion is less strict than the one for parameter estimators, because population values for standard errors are unknown for finite samples and have to be estimated, introducing additional uncertainty (see Hoogland, 1999, p. 30).

The coverage rate of the 95%-confidence interval of an estimator is considered sufficient when, over replications, the proportion of intervals containing the population parameter is between 0.90 and 0.98. These asymmetric margins were chosen at face value to allow for some but not a lot of deviation.

The RMSE of an estimator is the square root of the sum of its squared bias and variance. Accuracy and precision are thus both taken into account in this statistic. When the RMSE of an estimator is small, the probability that the estimate differs much from the parameter value in any replication is small (Lindgren, 1993, p. 254). Among unbiased estimators, the most efficient estimator has the smallest RMSE. In the evaluation of our results, we use the RMSE as supplementary information to the relative bias, comparing the quality of estimators of different models for the various cells in the design, estimators having the smallest RMSE usually being preferred. Consequently, because of its comparative use, no absolute criterion is formulated for the RMSE.

### Model Fit

Model fit indices are examined to assess the estimated fit of the model to the sample data. In our study, the structure of the model underlying the generated data always

equals the structure of the proposed model. Therefore, in terms of item/LV relations, the tested model always fits the data. Lack of model fit can thus only be the result of random sample variations or the imposed violations of model assumptions; in particular, the distributional properties of the LV and the item variables.

In case of normal LV and item distributions, it is expected that the fit indices indicate a good fit. When model assumptions are violated, the robustness of the fit indices becomes apparent, as estimators may be biased as a result of these violations. Since in practice, fit indices are used to evaluate whether a hypothesized latent structure serves as a good approximation to the empirical relationships, fit indices are more useful when they are robust against violations of model assumptions such as nonnormal LV or item distributions — the indices are not used to assess whether model assumptions hold.

Although an abundance of indices is available for the FA-lin and FA-poly models, our primary interest is in the $\chi^2$ statistic, the root mean squared error of approximation (RMSEA; Steiger & Lind, 1980; Steiger, 1990), and the standardized root mean residuals (SRMR; Jöreskog & Sörbom, 1981; Bentler, 1995), since it is generally recommended to report these measures (e.g., Boomsma, 2000; Hu & Bentler, 1998). In addition, they are commonly used in practice to evaluate the fit of a model, as found in Chapter 2.

A variety of $\chi^2$ statistics have been proposed in the literature (e.g., Browne, 1984; Hu & Bentler, 1999; Yuan & Bentler, 1997). For the three parametric models, we report the statistic proposed by Yuan and Bentler (1997), $\chi^2_{YB}$, as it has been shown to produce proper results for IRT-grm (Maydeu-Olivares et al., 2011).

Since MPLUS does not generate the $\chi^2_{YB}$ statistic, we computed it using our own R code. As the direct computation of the derivatives of the model-implied covariance matrix to the estimated parameters, involved in the calculation of $\chi^2_{YB}$, is very time consuming, we use the approximation formulas derived by Maydeu-Olivares et al. (2011), given in their appendix. To verify our implementation of these formulas, we also computed the derivatives directly for a number of data sets. The results of this verification were good, justifying the use of the approximation formulas. The R code used to compute the $\chi^2_{YB}$ statistic is available from the author.

The sampling distributions of the $\chi^2_{YB}$ statistic and the RMSEA are compared to their theoretical distributions by means of visual inspection of quantile-quantile (Q-Q) plots. The sampling distribution is considered sufficiently equal to the theoretical distribution when the points of the Q-Q plot are on the $x = y$-line, indicating equality of the sample and the theoretical quantiles.

The RMSEA (e.g., F. Chen et al., 2008),

$$\text{RMSEA} = \sqrt{\frac{\text{NCP}}{df(n-1)}}, \tag{4.23}$$

with

$$\text{NCP} = \max\left(0, \chi^2 - df\right), \tag{4.24}$$

follows a rescaled noncentral $\chi^2$ distribution, where NCP is the sample estimate of the noncentrality parameter, *df* denotes the degrees of freedom, and $n$ is the sample size. We take values smaller than 0.06 to be indicative of a good fit, as was recommended by Hu and Bentler (1999), and is usually done in practice (see Chapter 2).

The SRMR is defined as (Jöreskog & Sörbom, 1986)

$$\text{SRMR} = \sqrt{\frac{1}{I(I+1)/2}\sum_{i=1}^{I}\sum_{j=1}^{i}\left[\frac{s_{ij}-\hat{\sigma}_{ij}}{\sqrt{s_{ii}}\sqrt{s_{jj}}}\right]^2}. \tag{4.25}$$

As the SRMR involves a comparison of the model-implied covariances with the sample covariances, the model-implied covariance matrix is required for the three parametric models. For FA-lin this matrix was already given in Equation 1.6 as:

$$\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}' + \boldsymbol{\Psi},$$

which in the unidimensional case becomes

$$\boldsymbol{\Sigma} = \boldsymbol{\lambda}\boldsymbol{\lambda}' + \boldsymbol{\psi}. \tag{4.26}$$

For IRT-grm $\boldsymbol{\Sigma}$ is calculated using the formulas from Maydeu-Olivares et al. (2011). For FA-poly these formulas are applied making use of the normal-ogive instead of the logistic function.

Analogous to the evaluation of the RMSEA, we take values smaller than the commonly applied performance criterion of 0.08 for the SRMR (Hu & Bentler, 1999) to be indicative of acceptable model fit.

For IRT-mok, Loevinger's $H_{scale}$ is taken as an indicator of model fit and compared to the performance criterion of 0.30 (see Sijtsma & Molenaar, 2002), with lower values indicating lack of fit, and higher values indicating a homogeneous scale.

### Latent Variable Scores

In practice, scales are often used for the sole purpose of estimating a respondent's LV score. It is therefore important for a model to produce LV score estimates that are close to the respondents' true LV scores. The various models produce LV scores that differ in the weights given to each item of the scale: FA-lin weights by the estimated loadings, FA-poly/IRT-grm take both the loadings/discriminations and the estimated thresholds/step-difficulties into account. IRT-mok uses unweighted sum scores.

In our evaluation of LV results, we are interested in two aspects. First, does the *shape* of the estimated distribution resemble the shape of the true distribution? To assess this, we focus on the difference in skewness of the estimated and sampled LV distributions. Second, does the *ordering* of respondents based on their estimated LV scores resemble the ordering based on their true LV scores? To this end, the association between the estimated and the true LV scores is investigated by inspecting scatterplots of the true and estimated LV scores. Kendall's (1938) rank correlation $\tau_a$ is estimated to quantify that association.

**Analysis of Variance**

In summary, model estimation is evaluated by examining the following performance variables:

1. Parameter estimators (plain or relative bias, RMSE, coverage rate),
2. Standard error estimators (relative bias, root mean square error),
3. Model fit ($\chi^2_{YB}$, RMSEA, SRMR),
4. Latent variable score estimators (skewness, Kendall's $\tau_a$).

We use a multivariate or univariate analysis of variance (ANOVA) to determine which of the explanatory variables markedly affect the response variables (cf. Finch, 2010). Since the sample size for these analyses equals the number of replications generated per performance variable and $R = 1000$ is very large, each ANOVA is over-powered, i.e., it will detect even the very small effects. Hence, in the evaluation of our results, we emphasize the effect size of (combinations of) the explanatory variables, focusing on design factors that explain a substantial proportion of the variance of our performance variables. To this end, we report $\eta^2$ for the univariate ANOVAs, which is defined as (e.g., Cohen, 1973)

$$\eta^2 = \frac{SS_f}{SS_t}, \tag{4.27}$$

where $SS_f$ denotes the sum of squares of effect $f$ and $SS_t$ the total sum of squares.

Although there has been some debate about whether to use $\eta^2$ or partial $\eta^2$, $\eta^2_p$ (e.g., Cohen, 1973; Levine & Hullett, 2002; Olejnik & Algina, 2000; Tabachnick & Fidell, 1983), in which $SS_t$ is replaced by the sum of $SS_f$ and the error sum of squares $SS_e$, thus partialling out the variance accounted for by the other design factors, we chose $\eta^2$ for our univariate ANOVAs for its ease of interpretion: the proportion of variance of the response variable accounted for by the effect. In the evaluation of $\eta^2$, we keep in mind Cohen's (1988) guidelines of interpreting $\eta^2$ values of 0.02, 0.13, and 0.26 as small, medium, and large, respectively.

For the MANOVAs, we report the multivariate partial $\eta^2_p$, defined as (e.g., Kline, 2004)

$$\eta^2_p = 1 - \Lambda^{1/s}, \tag{4.28}$$

with

$$\Lambda = |\mathbf{W}|/|\mathbf{T}| \tag{4.29}$$

and

$$s = min(df_f, k) \tag{4.30}$$

where $\Lambda$ denotes Wilk's lambda, $|\mathbf{W}|$ is the determinant of the within-groups matrix $\mathbf{X}'\mathbf{X} - \bar{\mathbf{X}}'\bar{\mathbf{X}}$, which is also known as the sums of squares and cross-products (SSCP) matrix, $\mathbf{T}$ is the total SSCP matrix, $df_f$ denotes the degrees of freedom for effect $f$,

and $k$ is the number of response variables. In contrast to the univariate $\eta_p^2$, which is *larger* than $\eta^2$ in case of multiple explanatory variables, in the multivariate case, $\eta_p^2$ is *smaller* than $\eta^2$ when $s > 1$ and thus a more conservative estimate of effect size. Furthermore, the interpretational advantage of $\eta^2$ over $\eta_p^2$ does not hold in the multivariate case, as neither $\eta^2$ nor $\eta_p^2$ equals the proportion of explained variance of the response variables accounted for by the effect. Hence, we report the more conservative $\eta_p^2$ for the MANOVAs.

To apply a (multivariate) analysis of variance ((M)ANOVA), three conditions have to be met: (a) independence of the observations, (b) (multivariate) normality of the response variable(s), and (c) equality of the population (co)variances (e.g., Stevens, 2002, p. 176). First of all, the independence of the observations is assured by the pseudo-random sampler used in the data generation process and the ensuing analysis. The second assumption is of special interest, as the distribution of $\chi^2$-based fit measures are not expected to be normal. Fortunately, the results of a univariate ANOVA have been shown to be quite robust to violations of this assumption (Stevens, 2002, p. 257, p. 262ff.). Furthermore, violations of the second and third assumption result in an improper Type I error. Since we are not much concerned with the Type I error, due to the large sample size, but use the ANOVAs merely as an aid to focus on the most influential explanatory variables of our design, using effect size estimates, we decided to apply the (M)ANOVAs even to our intrinsically skewed response variables.

## 4.3 Expectations

From the previous chapter it is clear that the robustness of FA and IRT models has already been subject to a lot of research. The expectations presented in this section partly serve to assess the replicability of those previous findings. In our Monte Carlo study, we can combine the results from the separate fields of FA and IRT, and contribute by systematically comparing the models' robustness against combinations of LV and item skewness.

Since simulation studies on the properties of IRT-mok are scarce, we only formulate a few high-level expectations regarding this model. With our study we take a first step in evaluating the performance of IRT-mok under conditions of LV and item skewness in itself, and in comparison to the parametric models.

Recall from the previous chapter the factors of our design, i.e., the explanatory variables in the (M)ANOVAs:

- Estimation model (FA-lin, FA-poly, IRT-grm, IRT-mok)
- LV distribution (normal, right skew-normal)
- Scale shape (various combinations of normal, right-skewed, left-skewed, and bimodal item response distributions)
- Scale strength (strong: all items load strongly [0.80] on the LV; or mixed: 4 items strong [0.80], 4 medium [0.50], and 4 weak [0.30])

- Sample size (small: $n = 200$; medium: $n = 600$)

Here, we further substantiate the hypotheses for the simulation study, following the general expectations in Chapter 3. In composing our expectations we had to make some — unwarranted — generalizations. To enhance the transparency of our reasoning, we mention them explicitly:

- We generalized results for the discrimination parameter to loading parameters, and formulated our hypotheses (mostly) in terms of the loading parameter.
- We generalized results for the difficulty parameter to threshold parameters, and formulated our hypotheses in terms of the threshold parameter.
- In case of designs with dichotomous items including multiple difficulty levels, we took easy and difficult items to represent left- and right-skewed polytomous items, respectively.

We divide our expectations into five groups: expectations comparing the estimation models, and expectations concerning each of the estimation models, FA-lin, FA-poly, IRT-grm, and IRT-mok, separately.

1. Expectations for model comparisons
   (a) Under conditions of normal LV and item distributions, we expect FA-poly and IRT-grm to perform equally well with no parameter estimation bias (Forero & Maydeu-Olivares, 2009).
   (b) Of all estimation models, FA-lin is generally expected to perform worst (Coenders et al., 1997), as its assumption of a linear relationship between the LV and the observed variables is violated when items are categorical.
   (c) One exception to 1b is expected: Under the condition of congruent LV and item skewness, FA-lin could well outperform FA-poly, which has been found to produce biased parameters in case of a skewed LV (Flora & Curran, 2004; Rhemtulla et al., 2012).
   (d) IRT-grm is expected to outperform FA-poly slightly in case of a right skew-normal LV (Boulet, 1996; DeMars, 2010; Finch, 2010; Kay, 2004).
   (e) In case of item skewness, IRT-grm will outperform FA-poly, when combined with other unfavorable conditions such as small loadings and a small sample size (Forero & Maydeu-Olivares, 2009).
   (f) When the assumptions of the parametric models are met, they are presumed to be more powerful and provide more information than the nonparametric IRT-mok, but when they are violated, IRT-mok will be a good alternative and might even provide more useful results.
   (g) With regard to model fit, $\chi^2$ values are expected to be overestimated under conditions of nonnormality (Flora & Curran, 2004; Trierweiler, 2009), but

more so for FA-lin than for FA-poly (Rhemtulla et al., 2012). The $\chi^2_{YB}$ statistic, however, is tentatively expected to be quite robust to deviations from normality for both FA-lin and IRT-grm, as its performance under conditions of normality has been found quite good compared to other $\chi^2$ statistics (Maydeu-Olivares et al., 2011).

(h) LV score estimation is expected to be better for FA-poly and IRT-grm than for FA-lin and IRT-mok, especially in the tails of the distribution (Dumenci & Achenbach, 2008).

(i) Over all conditions and estimation models, we expect better results for the medium than for the small sample size.

2. Expectations for FA-lin

(a) When both the LV and item distributions are normal, FA-lin *parameter* estimators are expected to be negatively biased for both our small and medium sample size (Babakus et al., 1987; Beauducel & Herzberg, 2006; Coenders et al., 1997; DiStefano, 2002; Dolan, 1994; Rhemtulla et al., 2012; Trierweiler, 2009).

(b) When item variables are skewed but the LV distribution is normal, FA-lin parameter estimators are expected to be biased more severely than they are for normal items (Babakus et al., 1987; Boomsma, 1983; Coenders et al., 1997; DiStefano, 2002; Rhemtulla et al., 2012).

(c) When both the LV and item distributions are skewed in the same direction, FA-lin parameter estimators are expected to be biased about as much as in the normal condition, because the assumption of a linear relation between the LV and the item variables is not severely violated then (Coenders et al., 1997; Rhemtulla et al., 2012).

(d) Although the condition of a skew-normal LV distribution and normal item variables has not been investigated yet, we expect FA-lin parameter estimators to be biased, because the thresholds are not evenly spaced, hence violating the assumption of a linear relation between the LV and the item variables (cf. Coenders et al., 1997).

(e) FA-lin *standard error* estimators are of special interest, as the literature is inconclusive on this subject. We tentatively expect a negative bias of standard error estimators when sample size is small (Babakus et al., 1987; Rhemtulla et al., 2012), and under conditions of nonnormality (Babakus et al., 1987; DiStefano, 2002; B. O. Muthén & Kaplan, 1985; Rhemtulla et al., 2012).

(f) FA-lin *model fit* indices are expected to behave properly when item distributions are normal (Beauducel & Herzberg, 2006; Rhemtulla et al., 2012; Trierweiler, 2009).

(g) Under conditions of normality FA-lin $\chi^2_{YB}$ is tentatively expected to perform well, as (Maydeu-Olivares et al., 2011) found such behavior in case of multivariate normal (and thus continuous) items.

(h) FA-lin $\chi^2_{YB}$ values are expected to indicate underfit in case of skewed item variables, as this was found for other $\chi^2$ statistics (Boomsma, 1983; DiStefano, 2002; B. O. Muthén & Kaplan, 1985; Rhemtulla et al., 2012) and increasingly so for an increasing number of skewed items included in a scale (DiStefano, 2002).

(i) The RMSEA for FA-lin is expected to be quite robust against nonnormal items (DiStefano, 2002).

(j) The SRMR is also expected to perform well in case of nonnormal items (DiStefano, 2002).

(k) FA-lin *LV score* estimates are expected to deviate from the population values, especially in the tails of the distribution (Dumenci & Achenbach, 2008).

3. Expectations for FA-poly

(a) For normal LV and item distributions, FA-poly loading *parameter* estimators are expected to be unbiased (Beauducel & Herzberg, 2006; Boulet, 1996; Forero et al., 2009; Kay, 2004; Rhemtulla et al., 2012; Trierweiler, 2009).

(b) Threshold parameters are expected to be unbiased in case of normal LV and item distributions (Boulet, 1996; Finger, 2001).

(c) When item variables are skewed but the LV distribution is normal, we expect FA-poly loading parameter estimators to be unbiased (Coenders et al., 1997; DiStefano, 2002; Flora & Curran, 2004; Forero et al., 2009; Rhemtulla et al., 2012; Yang-Wallentin et al., 2010).

(d) FA-poly threshold parameter estimators of skewed items loading on a skew-normal LV are also expected to be unbiased (Forero & Maydeu-Olivares, 2009).

(e) In case of a skew-normal LV distribution and normal item variables, we expect FA-poly loading parameter estimators to be biased, because estimation of the polychoric correlations depends on the assumption of a normal underlying LV, which is violated then.

(f) When both the LV and the item distribution are right-skewed, loading parameter estimators are expected to be positively biased (Boulet, 1996; DeMars, 2010).

(g) FA-poly loading parameter estimators are expected to be negatively biased for a skew-normal LV combined with oppositely skewed items (Boulet, 1996; DeMars, 2010).

(h) For normal items loading on a right skew-normal LV, the outer, or extreme, threshold and difficulty parameters are expected to be underestimated, and middle thresholds/difficulties are expected to be unbiased or only slightly overestimated (Boulet, 1996; DeMars, 2010).

(i) For a skew-normal LV distribution, threshold parameters of correspondingly skewed items are expected to be overestimated (Boulet, 1996; DeMars, 2010).

(j) Thresholds of left-skewed items loading on a right skew-normal LV are expected to be overestimated and to a greater extent than those of right-skewed items (Boulet, 1996; DeMars, 2010).

(k) FA-poly loading parameter estimators are expected to be less efficient for a skew-normal LV distribution than for a normal LV distribution, in case of corresponding item distributions (Finch, 2010).

(l) For normal LV and item distributions, FA-poly *standard errors* of loading parameters are expected to be biased negatively only slightly, and thus acceptably (Flora & Curran, 2004; Forero et al., 2009).

(m) In case of skewed items loading on a normal LV, we expect no substantial bias of loading standard error estimators (Forero et al., 2009). However, as Rhemtulla et al. (2012) did find a considerable negative standard error bias for FA-poly loadings, regardless of the LV or item distribution, Expectations 3l and 3m should be considered tentative.

(n) Threshold standard error estimators are expected to be unbiased for normal items loading on a normal LV (Forero & Maydeu-Olivares, 2009).

(o) For skewed items loading on a skew-normal LV, threshold standard errors are expected to be underestimated only slightly, and thus acceptably (Forero & Maydeu-Olivares, 2009).

(p) When the LV and item distributions are congruently skewed, loading standard error estimators are expected to be negatively biased more substantially (Flora & Curran, 2004) and increasingly so for increasingly large loading/discrimination parameter values (DeMars, 2010).

(q) Under conditions of normality, we expect FA-poly $\chi^2$ estimators of *model fit* to be unbiased (Rhemtulla et al., 2012; Trierweiler, 2009) or perhaps slightly overestimated for the small sample size (Flora & Curran, 2004), and we expect these results to generalize to the $\chi^2_{YB}$.

(r) In case of skewed items, $\chi^2$ values have been found to be overestimated (Flora & Curran, 2004; Trierweiler, 2009). The mean-and-variance adjusted $\chi^2$, however, appears to be rather robust to both LV and item skewness (Rhemtulla et al., 2012), and we expect to replicate these findings for the $\chi^2_{YB}$. When LV or item distributions deviate from normality, we tentatively expect the $\chi^2_{YB}$ to perform relatively well (Maydeu-Olivares et al., 2011).

(s) RMSEA and SRMR are expected to perform well for all LV and item distributions (Beauducel & Herzberg, 2006; Trierweiler, 2009).

(t) Regardless of the LV distribution, FA-poly *LV score* estimates are expected to be similar to the population values (Dumenci & Achenbach, 2008).

4. Expectations for IRT-grm

(a) For normal LV and item distributions, IRT-grm loading *parameter* estimators are expected to be slightly underestimated, but not substantially so (Forero & Maydeu-Olivares, 2009; Boulet, 1996; Stone, 1992).

(b) Loading/discrimination parameter estimation is expected to be more accurate for large than for small loading/discrimination values (Forero & Maydeu-Olivares, 2009). This expectation is tentative, as discrimination parameters have also been found to be unbiased for moderate and underestimated for large discrimination parameter values under conditions of normality (Boulet, 1996; Finger, 2001).

(c) Threshold parameters of normal items loading on a normal LV are expected to be unbiased in case of moderate values, in the range of about $[-1.5, 1.5]$ (Boulet, 1996; Finger, 2001).

(d) Such parameters are expected to be biased towards the extremes (negative bias for negative values and positive bias for positive values) in case of more extreme values (Boulet, 1996; Stone, 1992).

(e) When the LV distribution is normal and items are skewed, IRT-grm loading parameter estimators are expected to be unbiased (Forero & Maydeu-Olivares, 2009).

(f) Threshold parameter estimators are expected to be unbiased for skewed items loading on a normal LV (Forero & Maydeu-Olivares, 2009).

(g) In case of a right skew-normal LV distribution and corresponding item distributions, loading and discrimination parameters are expected to be overestimated either only slightly (Boulet, 1996) or substantially (DeMars, 2010; Stone, 1992).

(h) Loading parameters are expected to be substantially underestimated for left-skewed items loading on a right skew-normal LV (Boulet, 1996; DeMars, 2010).

(i) For normal items loading on a right skew-normal LV, the outer, or extreme, threshold and difficulty parameters are expected to be underestimated, and middle thresholds/difficulties are expected to be unbiased (Boulet, 1996; DeMars, 2010).

(j) For a skew-normal LV distribution, threshold and difficulty parameters of correspondingly skewed items are expected to be overestimated (Boulet, 1996; DeMars, 2010; Stone, 1992).

(k) Thresholds and difficulties of left-skewed items loading on a right skew-normal LV are expected to be overestimated and to a greater extent than those of right-skewed items (Boulet, 1996; DeMars, 2010).

(l) *Standard error* estimators of IRT-grm loading parameter estimators are expected to be unbiased when the LV and item distributions are normal (Forero & Maydeu-Olivares, 2009).

(m) A normal LV combined with skewed item variables is expected to result in unbiased loading standard error estimators for strong items (Forero & Maydeu-Olivares, 2009).

(n) Threshold standard error estimators are expected to be unbiased in case of normal items loading on a normal LV (Forero & Maydeu-Olivares, 2009).

(o) Skewed items loading on a normal LV are expected to results in a slight but acceptable underestimation of threshold standard errors (Forero & Maydeu-Olivares, 2009).

(p) For a skew-normal LV, IRT-grm loading standard error estimators are expected not to be substantially biased, regardless of the item distribution (DeMars, 2010).

(q) Unfortunately, Monte Carlo research including the evaluation of *model fit* for IRT-grm is mostly lacking. We expect the $\chi^2_{YB}$ to behave properly under conditions of normality (Maydeu-Olivares et al., 2011).

(r) When LV or item distributions deviate from normality, we tentatively expect the $\chi^2_{YB}$ to perform well.

(s) Although unprecedented, we expect RMSEA fit results based on the $\chi^2_{YB}$ to behave at least equally well as the $\chi^2_{YB}$ itself. Thus, we tentatively expect proper results for all LV and item distributions.

(t) Regardless of the LV or item distribution, we expect IRT-grm *LV score* estimates to be similar to the population values (Dumenci & Achenbach, 2008; Stone, 1992). LV scores estimates in the tails of the distribution are expected to be closer to the mean (Stone, 1992).

5. Expectations for IRT-mok

(a) For nonnormal LV distributions, IRT-mok is expected to perform relatively well, compared to the parametric models (Tate, 2003).

(b) IRT-mok *LV score* estimates, i.e., unweighted sum scores, are expected to deviate from the population values, especially in the tails of the distribution (Dumenci & Achenbach, 2008).

These expectations are summarized in Table 4.2. The main axes on which the expectations are presented here are item skewness (normal versus skewed) and LV distribution (normal versus right skew-normal), because they are the most prominent factors in our data configurations. Furthermore, scale strength is only varied in the

*Table 4.2.* Summary of expectations.

| | LV | Normal distribution | | | | Right skew-normal distribution | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | $\hat{\omega}$ | $\hat{se}(\hat{\omega})$ | Model | LV score | $\hat{\omega}$ | $\hat{se}(\hat{\omega})$ | Model | LV score |
| Item | Model | Bias | Bias | Fit | Bias | Bias | Bias | Fit | Bias |
| Normal | FA-lin | − | − | √ | − | ± | − | √ | − |
| | | 2a | 2e | 2f/2g | 2k | 2d | 2e | 2f | 2k |
| | FA-poly | √ | √ | √ | √ | ±/√/−[a] | √ | √ | √ |
| | | 3a/3b | 3l/3n | 3q/3s | 3t | 3e/3h | 3l | 3r/3s | 3t |
| | IRT-grm | √/−[a] | √ | √ | √ | √/−[b] | + | √ | √ |
| | | 4a/4b/4c/4d | 4l/4n | 4q/4s | 4t | 4i | 4p | 4r/4s | 4t |
| Skewed | FA-lin | − | − | −/√[d] | −/−[c] | − | − | −/√[d] | − |
| | | 2b | 2e | 2h/2i/2j | 2k | 2c | 2e | 2h/2i/2j | 2k |
| | FA-poly | √ | √ | √ | √ | +/−/+/+[e] | − | √ | √ |
| | | 3c/3d | 3m | 3r/3s | 3t | 3f/3g/3i/3j | 3p | 3r/3s | 3t |
| | IRT-grm | √ | √[f] | √ | √ | +/−/+ | √ | √ | √ |
| | | 4e/4f | 4m | 4r/4s | 4t | 4g/4h/4j | 4p | 4r/4s | 4t |

*Note.* √ indicates good expected performance; +, −, and ± indicate expected positive, negative, and unspecified bias, respectively. [a] Expectations differ for loading/discrimination, and inner and outer threshold parameters. [b] Expectations differ for inner and outer threshold parameters. [c] Bias for right-skewed item and LV distributions similar to the normal-normal case. [d] Good performance for RMSEA and SRMR. [e] For loading parameters of right-and left-skewed items, and threshold parameters of right-and left-skewed items, respectively. [f] For loading parameters of strong items.

normal part of our design. In the nonnormal data configurations it is held constant at strong so as to keep the study manageable and to focus on the effects of violations of distributional assumptions under the preferable condition of a strong scale.

**In the Following ...**

Data configurations with a normal LV and normal item variables are the topic of the next chapter. Samples of data are generated representing a strong scale, with all loading parameters $\lambda_i = 0.80$, or a mixed scale, with varying loading parameters of $\lambda_i \in \{0.30, 0.50, 0.80\}$. In addition, the sample size is either $n = 200$ or $n = 600$. Results of applying the four scaling models to the resulting conditions are presented in the following chapter and discussed in the light of (a subset of) the aforementioned expectations.

Subsequently, in Chapter 6, we focus on conditions of a nonnormal LV and nonnormal item variables. The remainder of the expectations is discussed then.

# Chapter 5

# Simulation Study: Normal Configurations

In the previous chapter, the setup of our simulation study, comparing factor analysis of the sample covariance matrix (FA-lin), factor analysis of the estimated polychoric correlation matrix (FA-poly), the graded response item response theory model (IRT-grm), and the nonparametric Mokken item response theory model (IRT-mok) was unfolded. Furthermore, expectations based on the literature discussed in Chapter 3 were presented.

The factors in our Monte Carlo design are:

- Estimation model: FA-lin by means of maximum likelihood (FA-lin-ML), FA-poly by means of mean-and-variance adjusted weighted least squares (FA-poly-WLSMV), the graded response model by means of robust maximum likelihood (IRT-grm-MLR), and the nonparametric Mokken item response theory model (IRT-mok)
- Latent variable (LV) distribution: normal and right skew-normal
- Scale shape: various combinations of normal, right-skewed, left-skewed, and bimodal item response distributions
- Scale strength: strong (all items load strongly [0.80] on the LV), and mixed (four items strong [0.80], four medium [0.50], and four weak [0.30])
- Sample size: small ($n = 200$), and medium ($n = 600$)

These variables are investigated in samples of data consisting of 12 five-category items loading on a single LV. In the present chapter, we focus on the normal conditions, i.e., with normal item distributions and a normal LV. Under these conditions, only the FA-lin-ML assumption of multivariate normal (hence continuous) item variables is violated.

In addition to evaluating our expectations with regard to normal data, we build a frame of reference for the interpretation of the results of the nonnormal data configurations, presented in Chapter 6.

In the following sections, we first elaborate on the configuration of the normal data conditions and the setup of the meta-analyses of results. Next, results are presented of applying the four estimation models to the simulated samples of data. Subsequently, the results are discussed and evaluated with respect to the expectations brought forth in the previous chapter. And finally, based on our findings some recommendations are presented.

## 5.1 Method

### 5.1.1 Four Normal Data Configurations

The specification of the normal data configurations is given in Table 5.1. The cell names identify their specifications in the order: scale shape, LV distribution, scale strength, and sample size. For example, Cell *nNS2* consists of *n*ormal items loading on a *N*ormal LV; all items load *S*trongly ($\lambda = 0.80$) on the LV and the sample size equals *2*00. Cell *nNM6* is also configured to have *n*ormal items loading on a *N*ormal LV, but item strength is *M*ixed, i.e., strong ($\lambda_i = 0.80$) for four items, medium ($\lambda_i = 0.50$) for four items, and weak ($\lambda_i = 0.30$) for four items. Sample size equals *6*00 there.

For each cell of the design (discussed in the present and the next chapter) the value used for seeding the pseudo-random number generator is listed in Table D.1 of Appendix D.1.

*Table 5.1.* Data configuration for four cells of the design, representing no violations of assumptions.

| Cell | Scale shape | LV distribution | Scale strength | Sample size |
|---|---|---|---|---|
| nNS2 | 12 × normal | normal | 12 × strong | 200 |
| nNS6 | 12 × normal | normal | 12 × strong | 600 |
| nNM2 | 12 × normal | normal | 4 × strong<br>4 × medium<br>4 × weak | 200 |
| nNM6 | 12 × normal | normal | 4 × strong<br>4 × medium<br>4 × weak | 600 |

### 5.1.2 ANOVA Setup

As was explained in Chapter 4, we analyze the results of our Monte Carlo simulation study using a (multivariate) analysis of variance ((M)ANOVA) approach. To analyze the effect of our design factors on the performance variables, we performed meta-analyses by means of (M)ANOVA. The response variables in the (M)ANOVAs involve estimators of parameters, standard errors, fit indices, and LV scores. The number of observations for the (M)ANOVAs $N$ equals the number of estimation models times the number of design cells times the number of replications, and varies with the response variable. For example, in the MANOVA on the loading parameters, $N$ equals $3 \times 2 \times 2 \times 1000 = 12000$, since there are three estimation models involved, two types of scale strength, two sample sizes, and 1000 replications. The explanatory variables are factors in our Monte Carlo design, i.e., estimation model, scale strength, and sample size.

In case of loading, discrimination, and scalability parameters, and corresponding standard errors, where multiple parameters are estimated simultaneously, we applied a repeated-measures MANOVA to the relative bias (RB) of the estimators of each of the parameters of interest. Since the RB is constituted by $(\hat{\omega}_r - \omega)/\omega$, it is this quantity for each parameter, that served as the response variable in the reported MANOVAs. The RB-constituents of the 12 item parameters served as the response variables, which were taken as repeated measures, as they are not estimated independently. We specified three between-subjects variables: estimation model, scale strength, and sample size. In addition, we identified a within-subjects variable, item group, indicating the grouping of the item loading values in case of the mixed scale (four weak, four medium, and four strong).

For the threshold and step-difficulty parameters, we applied a repeated-measures MANOVA to the plain bias (PB)-constituents of the parameter estimators and to the RB-constituents of the standard error estimators, with an additional within-subjects variable, threshold type, indicating whether the parameter is one of the outer (extreme) or one of the inner (middle) parameters. The latter distinction is used to examine differences in results for inner and outer threshold parameter estimators such as were addressed in, e.g., Item 3h.

To the model fit and LV results we applied two univariate ANOVAs. For the fit results, we chose the standardized root mean residuals (SRMR) and the root mean squared error of approximation (RMSEA) based on the $\chi^2_{YB}$ as response variables. The SRMR is distributed approximately normally over the 1000 replications, as can be concluded from Appendix D.2, where normal Q-Q plot are presented of the SRMR statistics.

The $\chi^2_{YB}$ is available for all three parametric models, but its theoretical distribution differs for FA-lin and FA-poly/IRT-grm. The FA-lin model involves 24 parameters (12 loadings and 12 error variances), leading to a $\chi^2$ distribution with 54 degrees of freedom. For FA-poly and IRT-grm 60 parameters are estimated (12 loadings/discriminations and 48 thresholds/step-difficulties), resulting in a $\chi^2$ distribution with 18 degrees of freedom. As the RMSEA is corrected for the degrees of freedom, it serves

well as the response variable for the ANOVA and is used instead of the $\chi^2_{YB}$ itself. In addition, an ANOVA is applied to the SRMR.

For the LV results, Kendall's $\tau_a$ served as the response variable. The distributional properties of the latter statistic over our 1000 replications did not hamper the ANOVA, since its distribution was approximately normal, as is evident from the Q-Q plots in Appendix D.3.

As was explained in Chapter 4, for the ANOVAs, we report $\eta^2$ of the effects, which equals the proportion of variance of the response variable accounted for by the effect. For the MANOVAs, the multivariate partial $\eta_p^2$ is reported, which is interpreted as the proportion of variance explained by the effect when partialling out the variance accounted for by the other explanatory factors.

Since the (M)ANOVAs are over-powered, we focus on the effect sizes. Therefore, only effects statistically significant at $\alpha = 0.01$ and of size $\eta_p^2 > 0.01$ or $\eta^2 > 0.01$ are reported. Interaction effects not meeting the aforementioned requirements are not listed in the tables. For readability, the names of the main effects not meeting the requirements are listed in the tables, but their effect sizes are not.

## 5.2 Results

In the current section, results are presented of running our simulation of four data configurations, as well as applying the scaling models to the resulting data. We start by a report of peculiarities encountered in the data generation process and the application of the scaling models. Next, we turn to an investigation of the distribution of estimates as an initial check before interpreting any of the results, calling for some caution in the interpretation of FA-lin parameter estimation results, as we will see.

Subsequently, results are laid out for parameter estimators, corresponding standard error estimators, model fit indices, IRT-mok coefficients, and LV scores.

For the sake of conciseness, not all results are displayed in the text; more detailed results can be found in Appendix D. In the tables throughout the chapter, results are reported at a precision of three decimal places, as is justified in Appendix D.5.

### 5.2.1 Peculiarities

In the data generating process for Cells nNS2, nNS6, nNM2, and nNM6, two replications initially contained empty cells in the univariate item frequency tables, i.e., one or more response categories of one or more items were not selected by any of the simulated respondents. Those replications (see Table 5.2) were discarded and replaced with a new run.

The model estimation procedures converged for all samples.

*Table 5.2.* Data configurations for which replications were discarded due to empty cells in the univariate item frequency tables.

| Cell | # Empty-cell replications |
|---|---|
| nNS2 | 1 |
| nNM2 | 1 |

### 5.2.2 Distribution of Estimates

For the parametric models, the parameter estimates and corresponding standard error estimates in each cell are explored by investigating the distribution of

$$z_{\hat{\omega}} = \frac{\hat{\omega} - \omega}{\hat{se}(\hat{\omega})}, \tag{5.1}$$

which is asymptotically expected to be standard normal. Confidence intervals based on parameter and standard error estimates are only accurate when $z_{\hat{\omega}}$ is standard normal. The distribution of $z_{\hat{\omega}}$ is discussed for the parameters in Cell nNS6. For the other three cells under investigation here, similar results were found.

Figure 5.1 gives the distribution of $z_{\hat{\lambda}_1}$ in Cell nNS6 for the parametric models. In addition, the mean ($\bar{x}$), standard deviation (SD), skewness ($\varsigma$), and excess kurtosis ($\kappa$) are given, which equal 0, 1, 0, and 0, respectively, for the standard normal distribution. The straight line in the Q-Q plots depicts a perfect correspondence between the theoretical and the observed distribution of $z_{\hat{\omega}}$.

For FA-poly and IRT-grm, the distribution does not show a clear deviation from standard normal. For FA-lin, a consistent negative bias of the parameter estimator is immediately obvious. Furthermore, the variance of $z_{\hat{\lambda}_1}$ is smaller than expected.

The distribution of threshold estimates $z_{\hat{\tau}_{1.1}}$ in Cell nNS6 are presented in Figure 5.2 for FA-poly and IRT-grm. For both models the empirical distribution is not clearly deviant from the standard normal distribution. The discrete appearance of the FA-poly estimates is a result of the WLSMV estimation procedure, with univariate estimation of the thresholds (L. K. Muthén, personal communication, October 10, 2009).

Figure 5.3 shows the distribution of $z_{\hat{\alpha}_1}$ for the discrimination parameter $\alpha_1$ in Cell nNS6 for FA-poly and IRT-grm. The shape of the distribution of $z_{\hat{\alpha}_1}$ is about normal for both models.

The distribution of estimated step-difficulty parameters $\beta_{1.1} = -2.056$ and $\beta_{1.2} = -0.804$ is depicted in Figures 5.4 and 5.5, respectively, for Cell nNS6 and FA-poly and IRT-grm. Equation 5.1 is not applied to this parameter, because its standard error is not easily derived, as $\beta^\circ$ is a function of the ratio of $\lambda^\circ$ and $\tau^\circ$ (cf. Equation 4.5); instead, the untransformed parameter values are plotted. The distribution of the step-difficulty parameter estimates appears to be approximately normal.

In summary, the distribution of the parameter estimates is deviant for the loading parameters estimated by FA-lin. These estimators are biased, therefore caution should

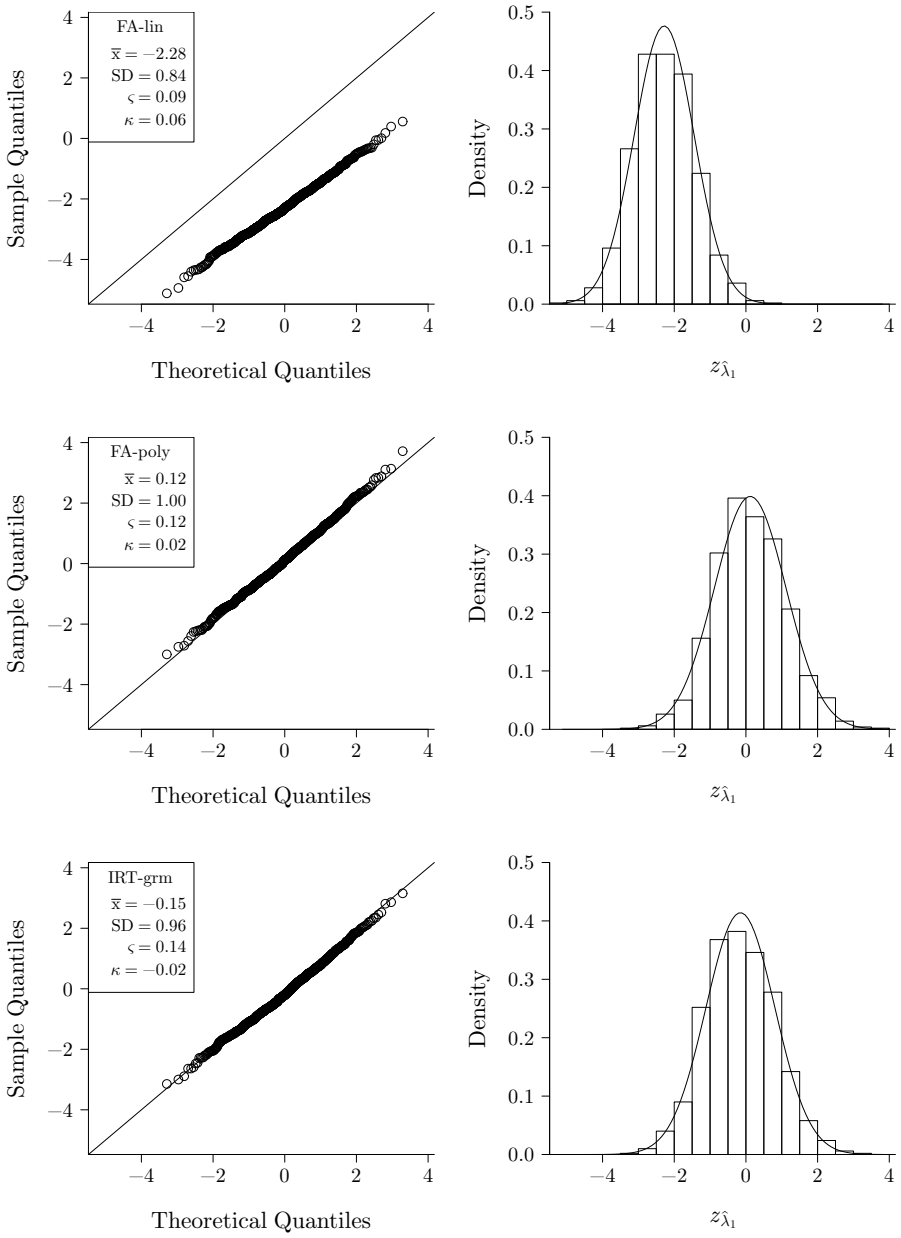



*Figure 5.1.* Q-Q plot and histogram of $z_{\hat{\lambda}_1}$ for FA-lin, FA-poly, and IRT-grm in Cell nNS6. Mean, standard deviation, skewness, and excess kurtosis are indicated in the inset. The curve in the histogram is a normal density curve with mean and variance taken from $z_{\hat{\lambda}_1}$. $n = 600$; $R = 1000$.

be taken in the interpretation of confidence intervals and corresponding coverage rates. The pattern of results found here approximately holds for all parameters in all cells.

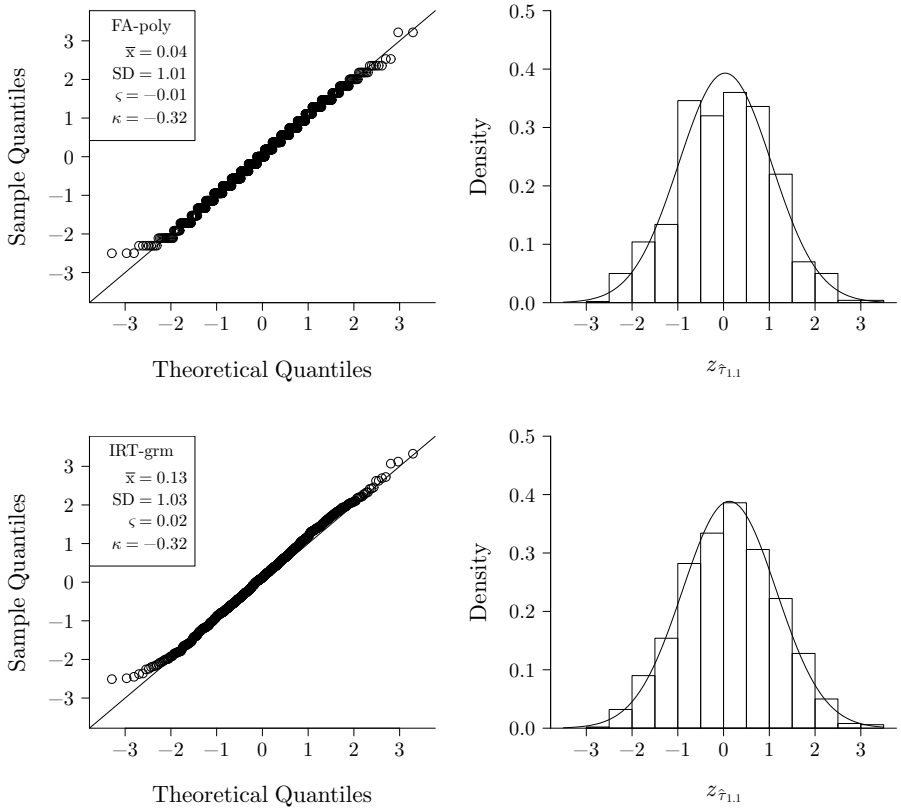



*Figure 5.2.* Q-Q plot and histogram of $z_{\hat{\tau}_{1.1}}$ for FA-poly and IRT-grm in Cell nNS6. $n = 600$; $R = 1000$.

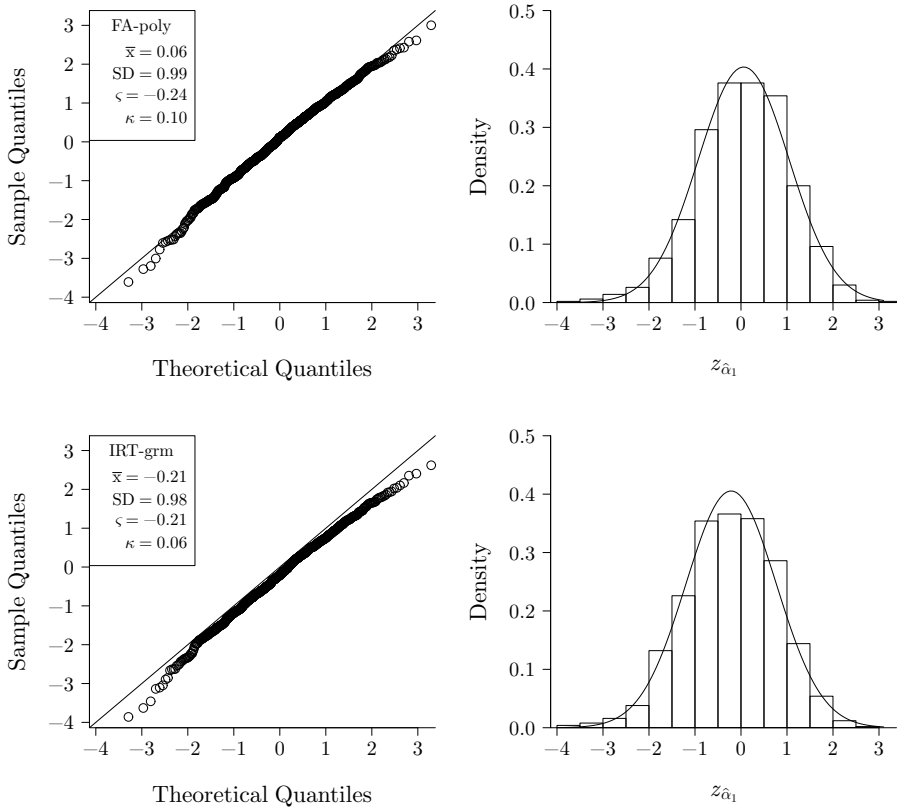



*Figure 5.3.* Q-Q plot and histogram of $z_{\hat{\alpha}_1}$ for FA-poly and IRT-grm in Cell nNS6. Mean, standard deviation, skewness, and excess kurtosis are indicated in the inset. The curve in the histogram is a normal density curve with mean and variance taken from $z_{\hat{\alpha}_1}$. $n = 600$; $R = 1000$.

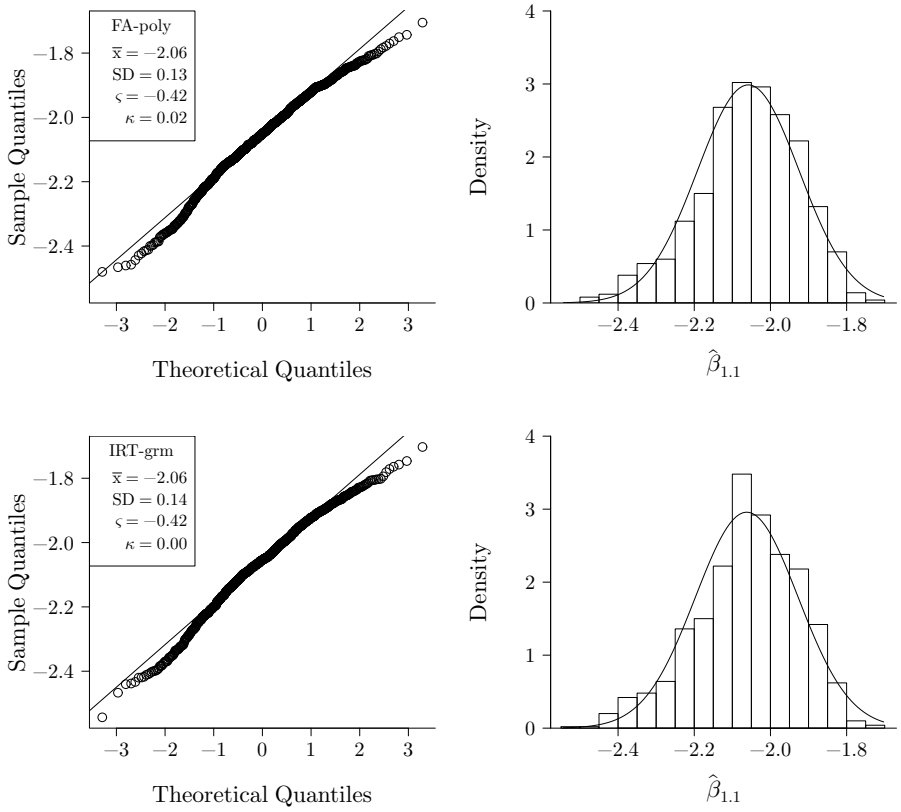



*Figure 5.4.* Q-Q plot and histogram of $\hat{\beta}_{1.1}$ for FA-poly and IRT-grm in Cell nNS6. $n = 600$; $R = 1000$.

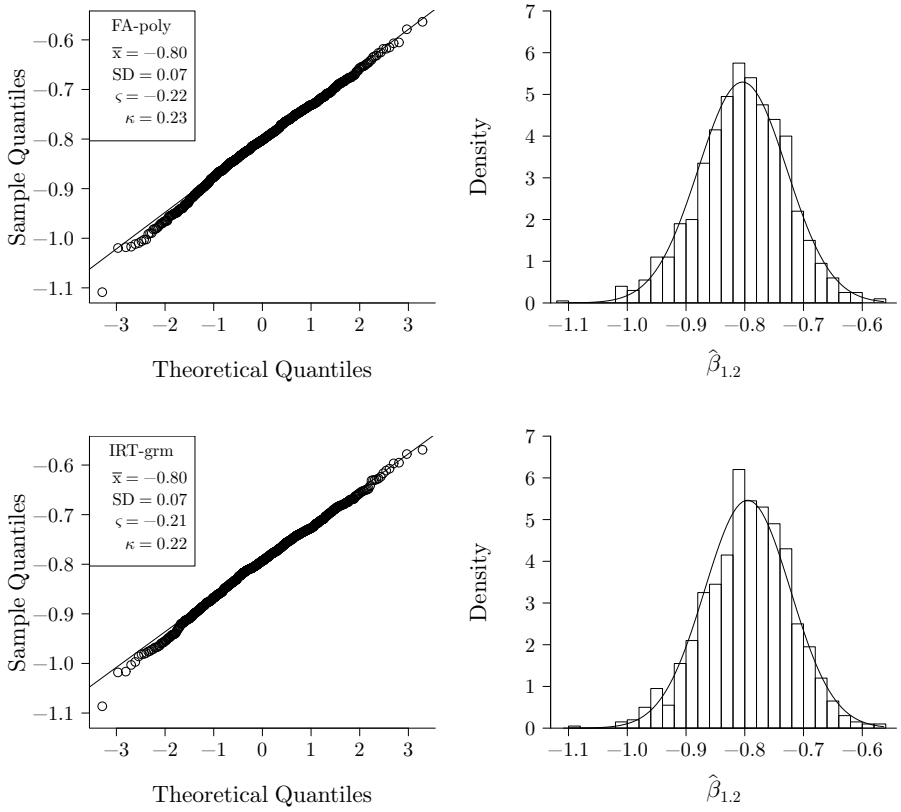



*Figure 5.5.* Q-Q plot and histogram of $\hat{\beta}_{1.2}$ for FA-poly and IRT-grm in Cell nNS6. $n = 600$; $R = 1000$.

### 5.2.3 Parameter and Standard Error Estimates

In the current section parameter and standard error estimation results are discussed. To investigate both the accuracy and precision of estimators, we examine the bias of estimators and the dispersion of estimates, respectively.

The starting point of each subsection is a MANOVA table providing the sizes of the effects of the explanatory variables of the Monte Carlo design quantified by $\eta_p^2$. Large effects are discussed with a preference of higher-order interaction effects over less complicated interaction effects or main effects, because a high-order interaction can cancel out the substantive interpretation of a lower-order effect. As mentioned before, the response variables in the MANOVAs are the RB-constituents $(\hat{\omega}_r - \omega)/\omega$, except for the threshold and step-difficulty parameter estimators. For thresholds and step-difficulties, we are interested in the plain, nonrelative deviations, because we consider a deviation from the population value important regardless of whether the threshold/step-difficulty is in the middle of the scale or at the end.

Next, we focus on the graphical display of the results, by means of boxplots of the estimates' deviation from the population value $(\hat{\omega}_r - \omega)$. We depict the PB- rather than the RB-constituents here, because the former also gives an indications of the estimator's precision.

To compare our results with the criteria set for the parameter and standard error estimators (5% and 10% deviation from the population value, respectively), the RB of parameter and standard error estimators is given in the text. Those values are averaged over the two sample sizes included in our design, unless they differ much between the sample sizes. In the tables included in Appendix D.4, results are listed separately for each sample size.

It should be noted that these three elements, MANOVA tables, PB-boxplots, and RB-values, are complementary: The MANOVA tables guide us towards the most important (combinations of) design factors affecting the model estimation performance; the PB-boxplots present the accuracy as well as the precision of estimators; and the RB-values provided in the text serve to evaluate the estimators' performance and to focus on the comparison of the results to the criteria set beforehand.

#### Loadings

**Parameters** Results of the MANOVAs applied to the RB-constituents of parameter and standard error estimates are presented in Table 5.3, listing $\eta_p^2 > 0.01$ for each effect with a $p$-value smaller than 0.01. The parameter estimation results are discussed first, and are illustrated by Figures 5.6 and 5.7, showing boxplots of the estimates' deviation from the population value $(\hat{\lambda}_r - \lambda)$ for the small and medium sample size, respectively. The mean deviation over $R = 1000$ replications, which equals the PB, is marked by a dot in each box. In line with convention, the median is marked by a horizontal line crossing each box. The boxplots are grouped by estimation model, and the individual boxes represent the results for a strong ($\lambda = 0.80$), medium ($\lambda = 0.50$),

*Table 5.3.* MANOVA results: $\eta_p^2$ per effect for RB of $\lambda$ parameters and of corresponding standard errors. $N = 12000$.

| Effect | Levels | Parameters $\eta_p^2$ | Standard errors $\eta_p^2$ |
|---|---|---|---|
| Model (m) | 3 | 0.335 | 0.105 |
| Scale strength (s) | 2 | 0.012 | |
| Sample size (n) | 2 | | 0.024 |
| m × s | 6 | 0.015 | |
| m × n | 6 | | 0.027 |
| Item group (ig) | 3 | 0.014 | 0.024 |
| m × ig | 9 | 0.012 | |
| s × ig | 6 | 0.014 | 0.023 |
| n × ig | 6 | | 0.122 |
| m × s × ig | 18 | 0.012 | |
| s × n × ig | 12 | | 0.051 |

*Note.* Listed effects are statistically significant at $\alpha = 0.01$ and are sized $\eta_p^2 > 0.01$.

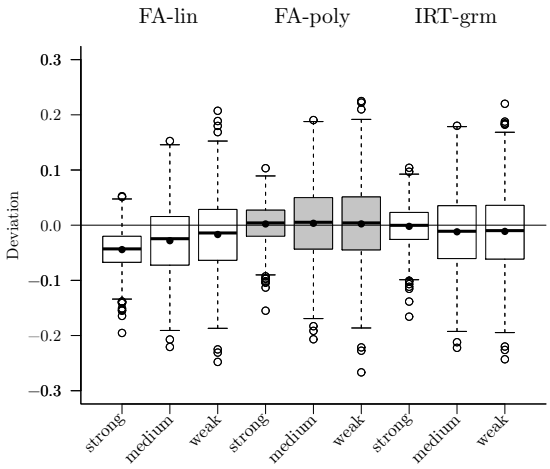



*Figure 5.6.* $\hat{\lambda}_i - \lambda_i$ for FA-lin, FA-poly, and IRT-grm. $n = 200$; $R = 1000$.

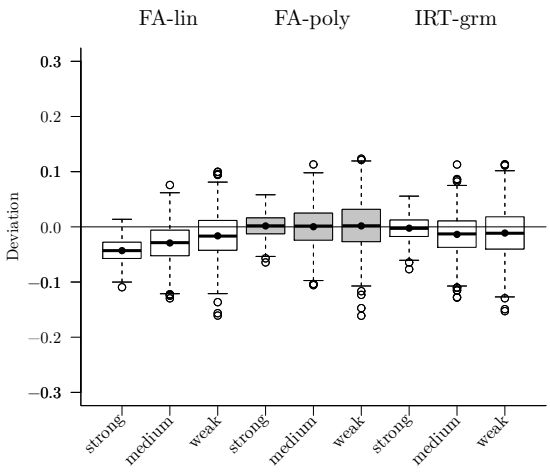



*Figure 5.7.* $\hat{\lambda}_i - \lambda_i$ for FA-lin, FA-poly, and IRT-grm. $n = 600$; $R = 1000$.

and weak ($\lambda = 0.30$) loading parameter. The horizontal line spanning the entire figure indicates the absence of bias.

From Table 5.3 we see that only the applied estimation model makes a substantial difference in the observed RB ($\eta_p^2 = 0.335$). There are additional significant effects, but these explain very little of the response variable's variance, the most meaningful of which is presumably the interaction effect of model, scale strength, and item group ($\eta_p^2 = 0.012$). The latter effect can be observed in Figures 5.6 and 5.7. The larger PB for the strong than the medium loading, which can be observed for FA-lin, is indicative of a rather stable RB. The rather constant PB of the medium and weak loading, which can be observed for IRT-grm, on the other hand, is indicative of an increasing RB with decreasing item strength.

FA-poly parameter estimators are unbiased and quite stable over different loading values (average RB of 0.3%; see Tables D.3, D.6, D.9 and D.12), which is in accordance with our Expectation 3a. FA-lin parameter estimators are consistently negatively biased with an average RB of −5.6%, −5.8%, and −6.4% for a strong, medium, and weak loading, respectively (see Tables D.2, D.5, D.8 and D.11), which is consistent with Expectation 2a. For IRT-grm, the RB increases more clearly as the population loading value decreases for both sample sizes, resulting in a substantial — but just acceptable — negative bias of the weak loading parameter (RB of −0.4%, −2.6%, and −4.3% for a strong, medium, and weak loading, respectively, averaged over both sample sizes; see Tables D.4, D.7, D.10 and D.13). This finding is in line with Expectations 4a and 4b.

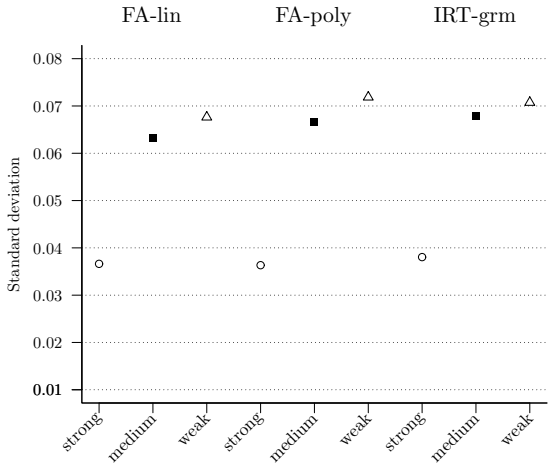



*Figure 5.8.* Standard deviation of $\hat{\lambda}_i$ for FA-lin, FA-poly, and IRT-grm. $n = 200$; $R = 1000$.

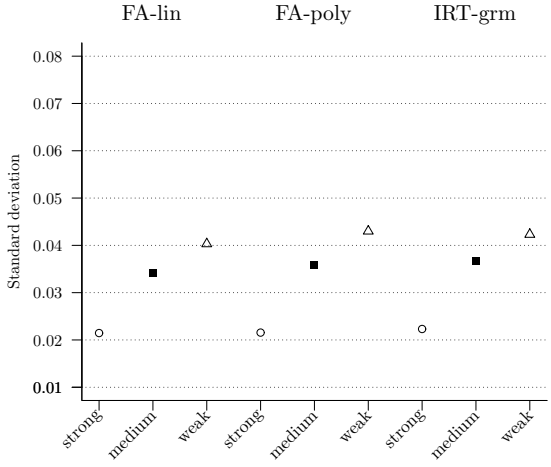



*Figure 5.9.* Standard deviation of $\hat{\lambda}_i$ for FA-lin, FA-poly, and IRT-grm. $n = 600$; $R = 1000$.

The precision of the loading parameter estimators can be observed in Figures 5.8 and 5.9, where the standard deviation of $\hat{\lambda}$ is plotted for the strong, medium, and weak item, as estimated by the three parametric models for the small and medium sample size, respectively. Comparing Figure 5.8 to Figure 5.6, we see that larger standard deviation values in the former plot are consistent with larger boxes and whiskers in the latter.

Comparing the two sample sizes, it can be noticed that $n = 600$ results in a smaller standard deviation of the estimates than $n = 200$. This holds for all parameter estimators and fits our expectation that a larger sample size results in better, i.e., generally more precise, parameter estimators. The bias, however, is not affected by sample size.

Comparing the models, we observe that for the small sample size, FA-lin is more precise than FA-poly and IRT-grm, especially in case of a medium or weak loading. However, the differences are small, and even smaller for the medium sample size. Furthermore, the standard deviation of loading estimates increases for decreasing loading values. Thus, for all parametric models, the precision of $\hat{\lambda}$ is greater for a larger sample size and larger loadings.

As the information conveyed in the standard error plots can also be retrieved from the boxplots — a larger standard deviation is reflected in a wider box and larger whiskers — they are not provided for the remaining parameters, the results of which are presented in the coming sections.

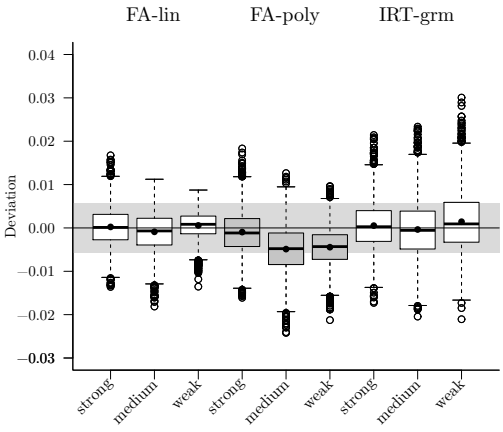



*Figure 5.10.* $\hat{se}(\hat{\lambda}_i) - sd(\hat{\lambda}_i)$ for FA-lin, FA-poly, and IRT-grm. The grey area represents an approximation of the margin of deviation around the true value considered acceptable. $n = 200$; $R = 4000$.

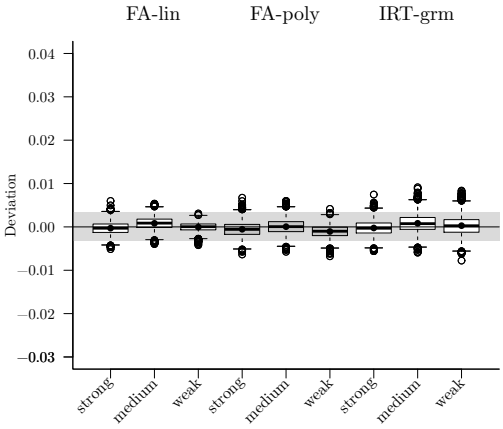



*Figure 5.11.* $\hat{se}(\hat{\lambda}_i) - sd(\hat{\lambda}_i)$ for FA-lin, FA-poly, and IRT-grm. The grey area represents an approximation of the margin of deviation around the true value considered acceptable. $n = 600$; $R = 4000$.

**Standard Errors** The results from the MANOVA on the RB of loading standard errors are presented in the last column of Table 5.3. A number of small effects are found, the largest of which are: the interaction of sample size and item group ($\eta_p^2 = 0.122$), the main effect for model ($\eta_p^2 = 0.105$), and the interaction of scale strength, sample size, and item group ($\eta_p^2 = 0.051$).

Figures 5.10 and 5.11 display the deviation of the loading standard errors from the parameter standard deviations $\hat{se}(\hat{\lambda}_r) - sd(\hat{\lambda})$ for the small and medium sample size, respectively, thus providing information of both the accuracy and precision of the standard error estimators. Because the bias of individual parameters of the same loading value varied considerably, the deviation of these standard error estimates are taken together in the boxplots, resulting in 4000 replications per boxplot, rather than 1000.

From these figures, we can observe the reported effects. For the small sample size, FA-poly stands out with underestimated standard errors (average RB of $-2.4\%$, $-7.5\%$, and $-6.0\%$ for strong, medium, and weak loading values, respectively; see Tables D.3, D.6, D.9 and D.12). These values, however, are all considered acceptable by our 10% criterion, and are in accordance with Expectation 3l. As we observe no standard error bias for FA-lin, our results are better than those reported by Rhemtulla et al. (2012), who found an RB of 8%–30% for sample sizes of 100 and 150, regardless of the LV or item distribution. Our results are more in line with their results for $n = 350$, for which they found no substantial standard error bias. Thus, our tentative expectation of biased FA-lin standard error estimators for the sample size of $n = 200$

(Expectation 2e) is not supported. This finding is further addressed in this chapter's discussion (Section 5.3). IRT-grm standard error estimators are, generally, unbiased (in line with Expectation 4l).

Furthermore, as the variance of standard error estimates is larger for the small sample size, we infer that $\hat{se}(\hat{\lambda})$ is more precise for a larger sample size, as would be expected.

### Thresholds

**Parameters** Table 5.4 provides the effect sizes of the MANOVAs we applied to the PB-constituents of the 48 threshold parameter estimators and RB-constituents of their corresponding standard error estimators. As threshold parameters are not part of the FA-lin model, the analyses were applied to FA-poly and IRT-grm results only, as reflected by the two levels of the explanatory variable model.

From Table 5.4 we infer that none of the explanatory variables of our Monte Carlo design have a meaningful effect on the PB of the threshold parameters. In Figures 5.12 and 5.13 we can observe that the smaller the item loading, the more the inner thresholds $\tau_{i2}$ and $\tau_{i3}$ are biased towards zero for IRT-grm.

For IRT-grm the PB of the inner thresholds of the medium and weak items is rather large, as the gross of the estimates is smaller in absolute value. This finding is contrary to Expectations 4c and 4d, stating that moderate threshold parameters are expected to be unbiased, whereas more extreme values are expected to be biased, respectively. These expectations were based on findings for step-difficulty parameters

*Table 5.4.* MANOVA results: $\eta_p^2$ per effect for PB of $\tau$ parameters and RB of corresponding standard errors. $N = 8000$.

| Effect | Levels | Parameters $\eta_p^2$ | Standard errors $\eta_p^2$ |
|---|---|---|---|
| Model (m) | 2 | | |
| Scale strength (s) | 2 | | |
| Sample size (n) | 2 | | 0.122 |
| m × n | 4 | | 0.012 |
| Item group (ig) | 3 | | 0.038 |
| s × ig | 6 | | 0.064 |
| n × ig | 6 | | 0.086 |
| s × n × ig | 12 | | 0.014 |
| Threshold type (t) | 2 | | 0.212 |
| s × t | 4 | | 0.012 |
| n × t | 4 | | 0.147 |
| s × n × t | 8 | | 0.083 |
| ig × t | 6 | | 0.271 |
| s × ig × t | 12 | | 0.052 |
| s × n × ig × t | 24 | | 0.012 |

*Note.* Listed effects are statistically significant at $\alpha = 0.01$ and are sized $\eta_p^2 > 0.01$.

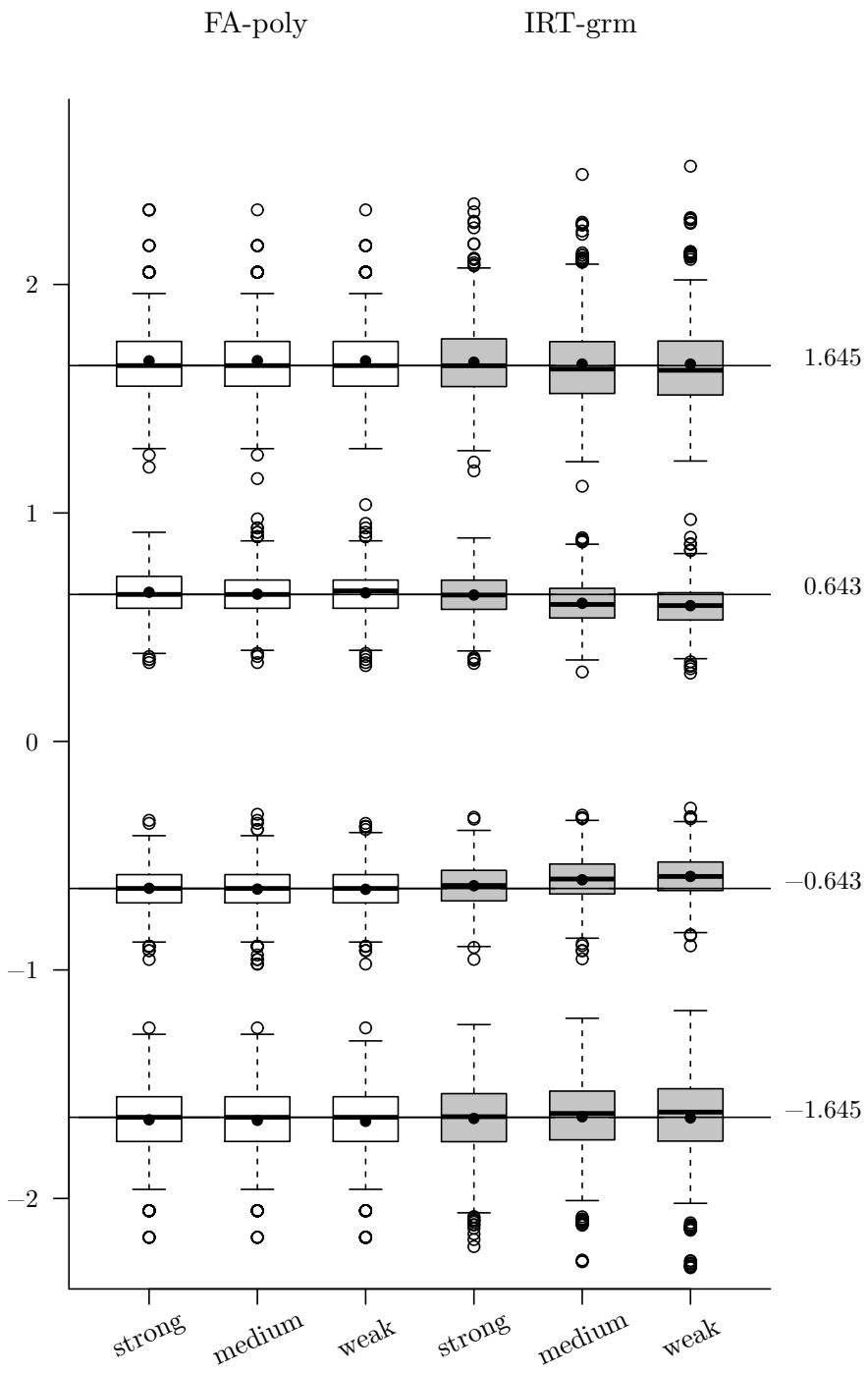



*Figure 5.12.* Parameter estimates $\hat{\tau}_{ic}$ for strong, medium, and weak items, as estimated by FA-poly and IRT-grm. The horizontal lines represent the true values, identified in the right margin. $n = 200$; $R = 1000$.

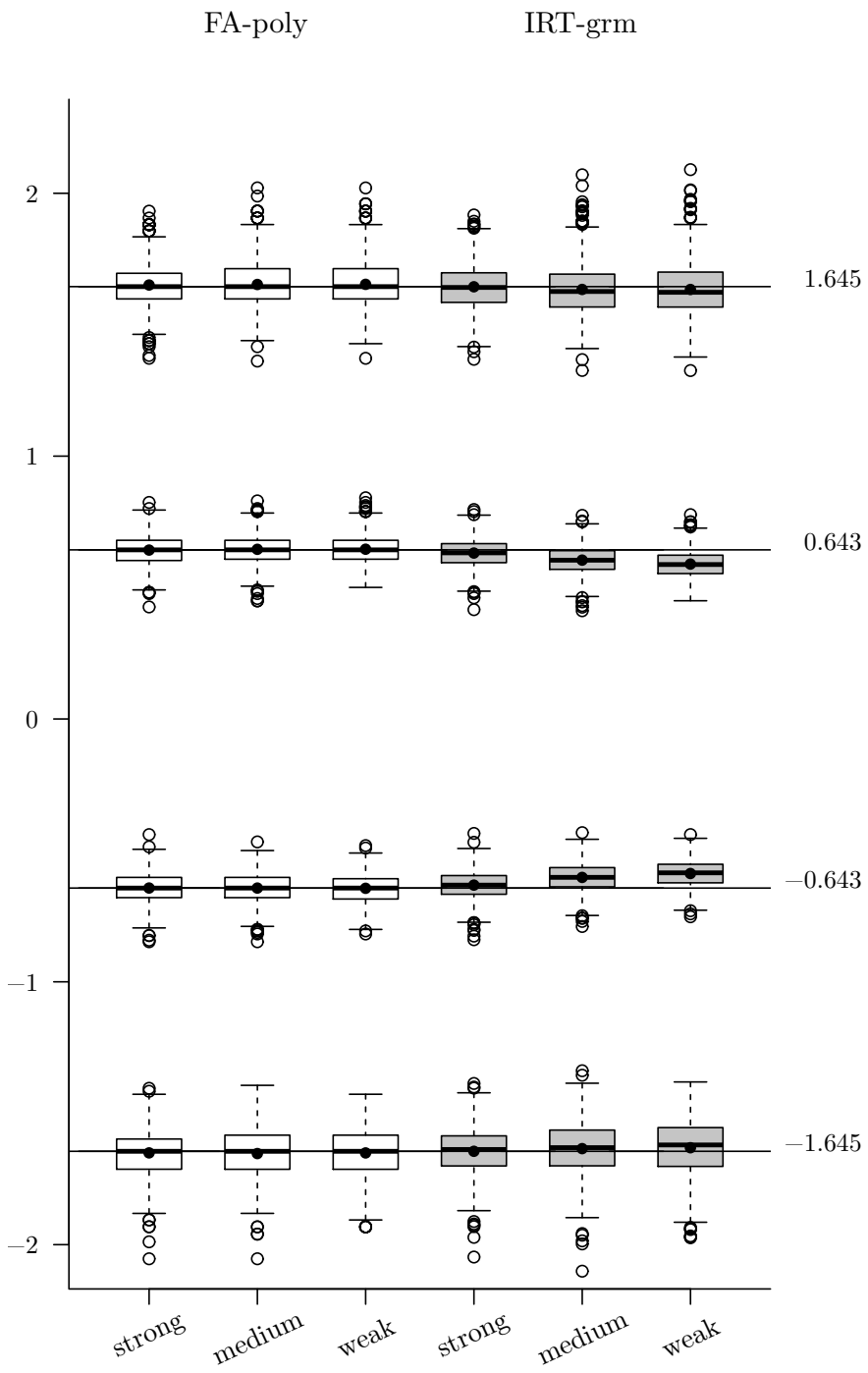



*Figure 5.13.* Parameter estimates $\hat{\tau}_{ic}$ for strong, medium, and weak items, as estimated by FA-poly and IRT-grm. The horizontal lines represent the true values, identified in the right margin. $n = 600$; $R = 1000$.

and the assumption of generalizability of item difficulty results to thresholds. As we will see in the paragraph on step-difficulty estimation results, only the generalizability assumption seems to be faulty. FA-poly threshold estimators are unbiased, which is in accordance with Expectation 3b.

Furthermore, the standard deviation of the estimates is greater for the outer thresholds than for the inner thresholds. This is also apparent from the root mean squared error (RMSE): Its value is consistently smaller for the inner thresholds than for the outer thresholds for both models (see Tables D.3, D.4, D.6, D.7, D.9, D.10, D.12 and D.13). This is probably due to the fact that estimation of parameters of more extreme values is less precise, because, generally, less information is available for such parameters, as most respondents are in the moderate answer categories, when items are normally distributed. Comparing the models, we can observe larger standard deviations for IRT-grm parameter estimates of the outer thresholds, especially in case of a medium or weak loading item. For the inner thresholds results are more similar for FA-poly and IRT-grm with a small advantage for IRT-grm.

**Standard Errors** The MANOVA results for the RB of the threshold standard error estimators are given in the last column of Table 5.4. As with the loading standard errors, there are many small significant effects. The largest effects are the interaction between item group and threshold type ($\eta_p^2 = 0.271$), the main effect of threshold type ($\eta_p^2 = 0.212$) and the interaction between sample size and threshold type ($\eta_p^2 = 0.147$). As there are no substantial effects of model, we conclude that FA-poly and IRT-grm threshold standard errors do not differ much. To graphically inspect these results we turn to Figures 5.14 and 5.15, where the deviation of the standard error estimates from the parameter standard deviation is shown for the four category thresholds (one per subplot) for small and medium sample size, respectively. As for the loading standard errors, estimates of each threshold belonging to an item of the same loading value are displayed together in one boxplot, resulting in 4000 replications per boxplot rather than 1000. Notice the different scales for the outer and inner threshold standard errors.

Comparing the grey areas in the upper and lower panels, representing an approximation to the 10%-margin of deviation around the true value considered acceptable, it is clear that the standard error estimators are more biased and less precise for the outer than for the inner thresholds, which is also apparent from the RMSE (see Tables D.3, D.4, D.6, D.7, D.9, D.10, D.12 and D.13). In addition, IRT-grm estimators seem to be slightly less precise than FA-poly estimators, as the dispersion of estimates is larger for IRT-grm. Notice the skewness of the distributions and the large number of outliers, both more prominent for IRT-grm than for FA-poly, indicating a tendency to occasionally seriously overestimate threshold standard errors.

Moreover, the estimators are far more precise for the medium sample size than for the small sample size. But most important, for both models and in each condition, the bias of all standard error estimators is within our 10% criterion, which is in accordance with Expectations 3n and 4n.

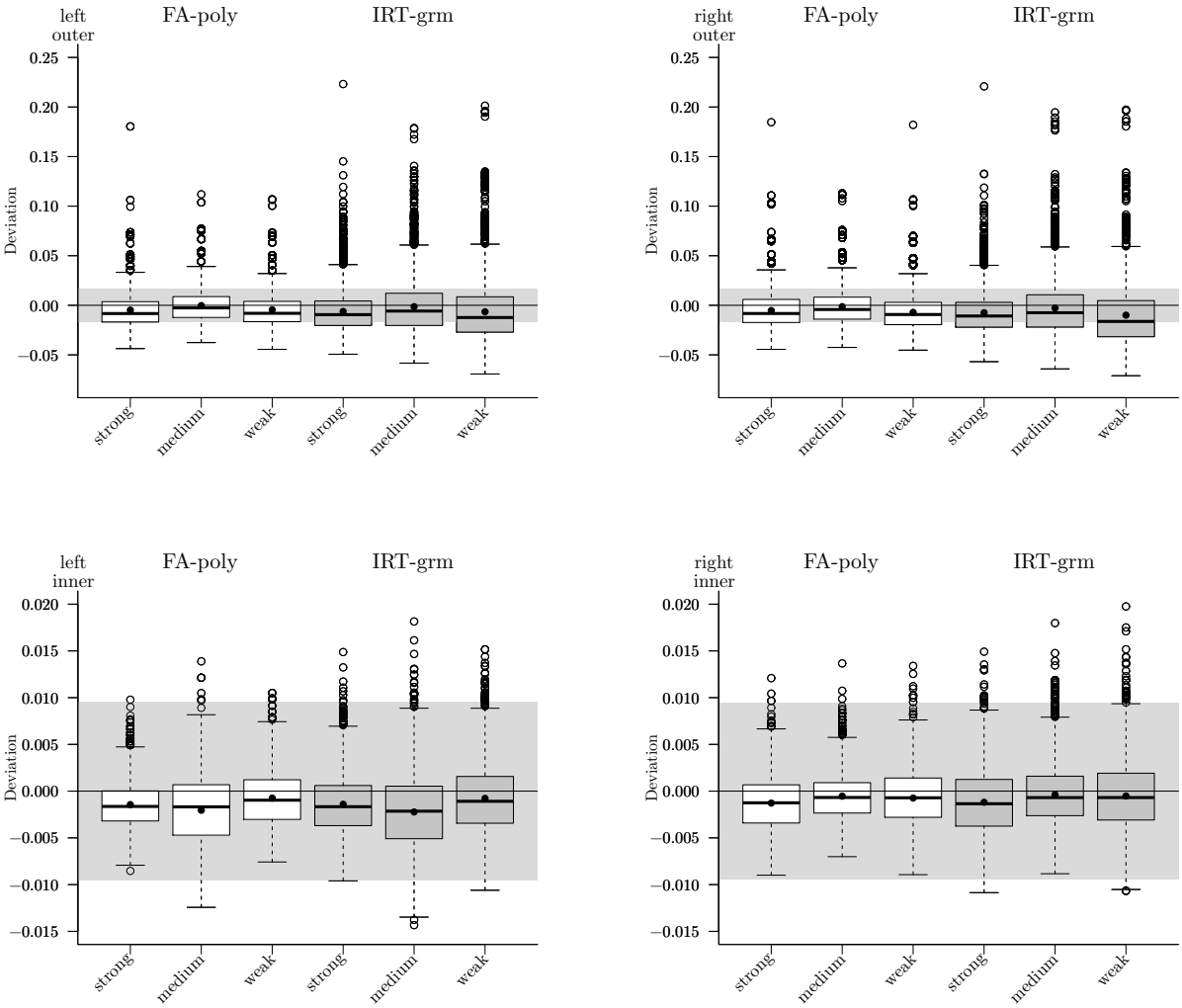



*Figure 5.14.* $\hat{se}(\hat{\tau}_{ic}) - sd(\hat{\tau}_{ic})$ for each of the four thresholds for FA-poly and IRT-grm. The grey area represents an approximation of the margin of deviation around the true values considered acceptable. $n = 200$; $R = 4000$.

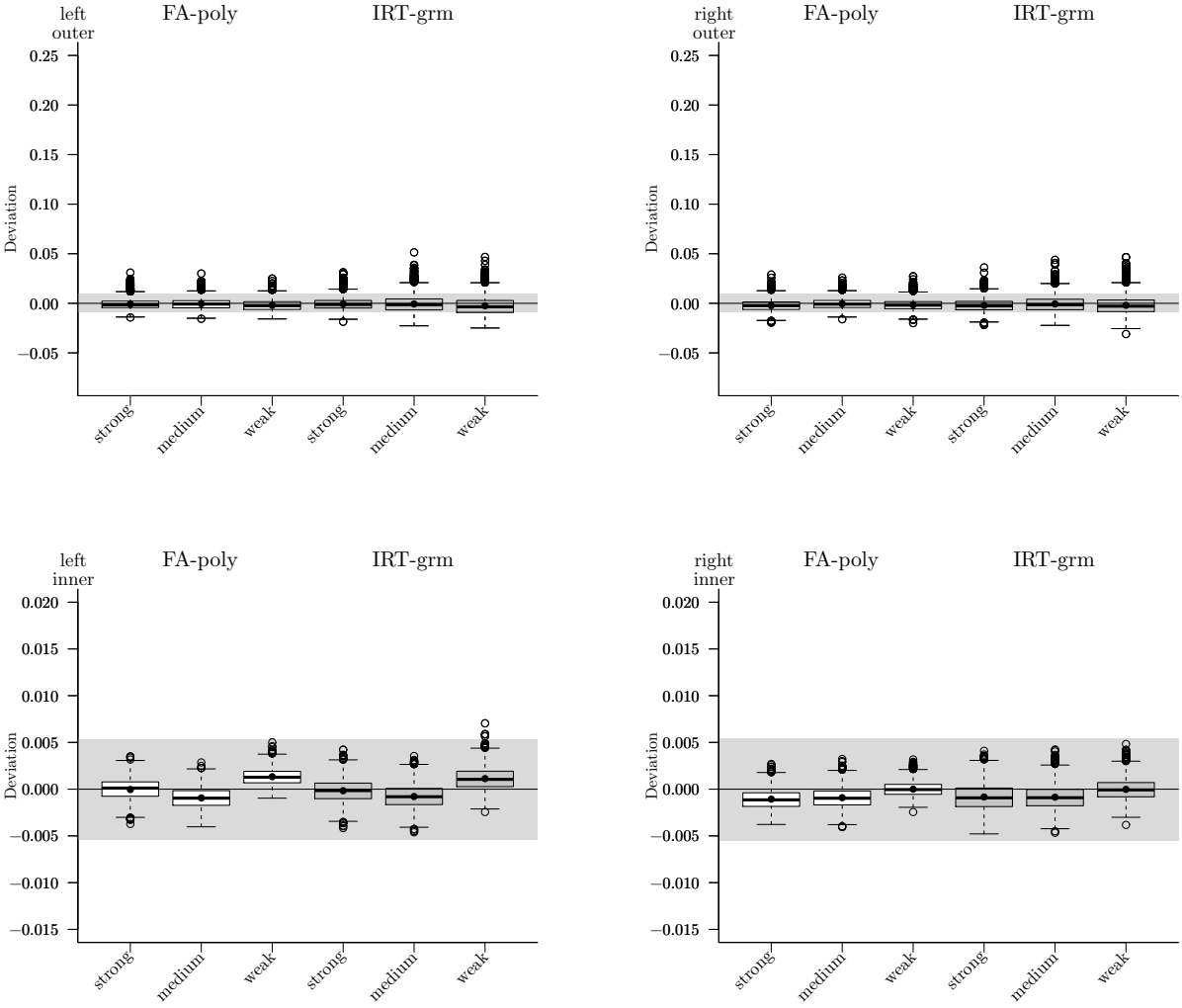



*Figure 5.15.* $\hat{se}(\hat{\tau}_{ic}) - sd(\hat{\tau}_{ic})$ for each of the four thresholds for FA-poly and IRT-grm. The grey area represents an approximation of the margin of deviation around the true values considered acceptable. $n = 600$; $R = 4000$.

### Discrimination and Step-Difficulty Parameters

As was explained in Chapter 4, data were generated under the FA-poly model, although we could equivalently have used an IRT parameterization. To facilitate the comparison with IRT research, we also present the parameter and standard error estimation results of the IRT discrimination $\alpha$ and step-difficulty $\beta$ parameters, albeit more briefly than the FA-parameterized results. We refer to Section 4.1.2 for the relations between the FA and IRT parameters.

**Discrimination Parameters** In Table 5.5 results are presented of the MANOVA on the RB of discrimination parameter estimators and their standard errors. The analysis was applied to FA-poly and IRT-grm results only, as reflected by the two levels of the explanatory variable model. For the parameters we find similar effects to those of the loading parameters. The largest effect is the main effect of model ($\eta_p^2 = 0.056$). It is smaller than the main effect of model in the MANOVA on loading parameter estimates ($\eta_p^2 = 0.335$), because the FA-lin results included in that analysis are not part of the analysis of discrimination parameters.

From Figures 5.16 and 5.17 we see that both FA-poly and IRT-grm discrimination parameters are, generally, unbiased, which is in line with Expectations 3a and 4a (worded in terms of loading parameters).

For IRT-grm we do find an increasing RB with a decreasing discrimination parameter value ($-0.05\%$, $-2.7\%$, and $-4.1\%$ RB for strong, medium, and weak items, respectively). This is in line with our tentative Expectation 4b, based on research by Forero and Maydeu-Olivares (2009), but contrary to findings of Boulet (1996) and Finger (2001), who reported an increase in bias with increasing item discrimination.

*Table 5.5.* MANOVA results: $\eta_p^2$ per effect for RB of $\alpha$ parameters and of corresponding standard errors. $N = 8000$.

| Effect | Levels | Parameters $\eta_p^2$ | Standard errors $\eta_p^2$ |
|---|---|---|---|
| Model (m) | 2 | 0.056 | 0.262 |
| Scale strength (s) | 2 | 0.011 | |
| Sample size (n) | 2 | | 0.139 |
| m × s | 4 | | 0.011 |
| m × n | 4 | | 0.070 |
| s × n | 4 | | 0.018 |
| Item group (ig) | 3 | 0.018 | 0.023 |
| m × ig | 6 | 0.011 | 0.014 |
| s × ig | 6 | 0.018 | 0.034 |
| n × ig | 6 | | 0.105 |
| m × s × ig | 12 | 0.011 | 0.013 |
| s × n × ig | 12 | | 0.058 |

*Note.* Listed effects are statistically significant at $\alpha = 0.01$ and are sized $\eta_p^2 > 0.01$.

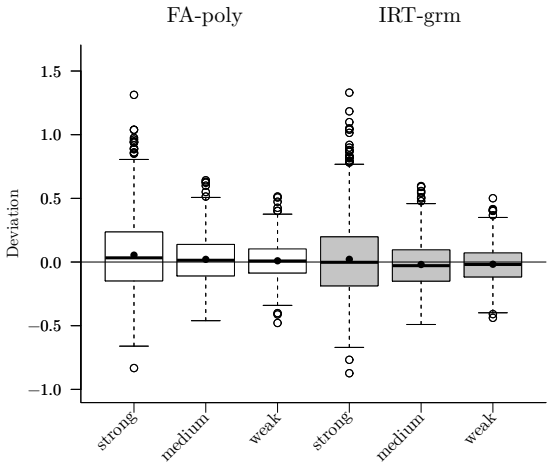



*Figure 5.16.* $\hat{\alpha}_i - \alpha_i$ for FA-poly and IRT-grm. $n = 200$; $R = 1000$.

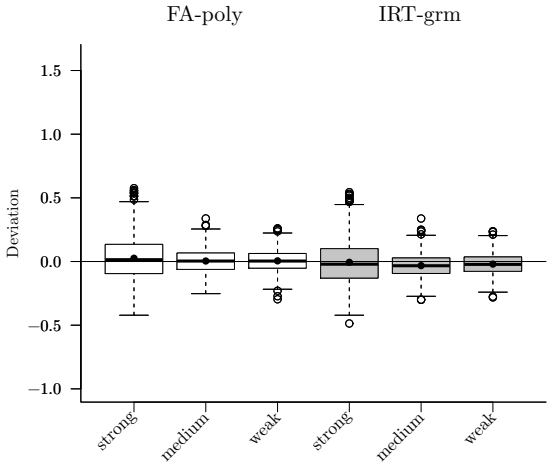



*Figure 5.17.* $\hat{\alpha}_i - \alpha_i$ for FA-poly and IRT-grm. $n = 600$; $R = 1000$.

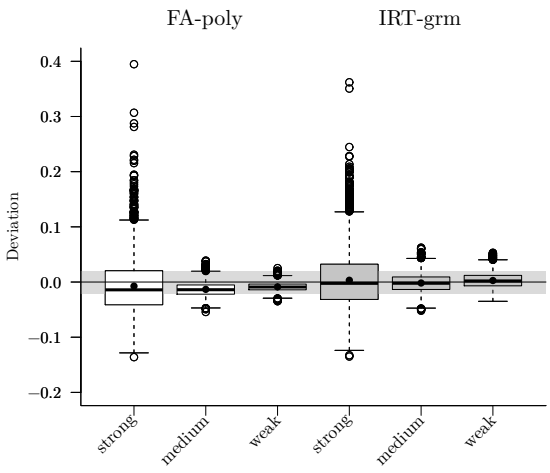



*Figure 5.18.* $\hat{se}(\hat{\alpha}_i) - sd(\hat{\alpha}_i)$ for FA-poly and IRT-grm. The grey area represents an approximation of the margin of deviation around the true value considered acceptable. $n = 200$; $R = 4000$.

The reason for these discordant findings is not clear. Perhaps the differences are due to the diverse software used in the studies, as Boulet (1996) used TESTFACT, Finger (2001) used his own personal code, and Forero and Maydeu-Olivares (2009) used MPLUS, as we did.

In addition, the estimators of both models are less precise for larger discrimination values and a smaller sample size.

**Discrimination Standard Errors** The last column of Table 5.5 contains the results of the MANOVA applied to the RB-constituents of standard error estimators. Once again, there are many significant but small effects. The largest effect is the main effect of model ($\eta_p^2 = 0.262$), signifying the underestimation of discrimination standard errors by FA-poly. However, with an RB of at most $-7.7\%$ (for a medium loading item, estimated by FA-poly for the small sample size) all deviations are within our 10% criterion, and thus considered acceptable.

Figures 5.18 and 5.19 graphically illustrate the results for a small and medium sample size, respectively. The graphs show the larger variance (also apparent from the RMSE; see Tables D.4, D.7, D.10 and D.13) and a skewed distribution with many outliers for standard error estimates of the largest discrimination parameter, i.e., for the strong item.

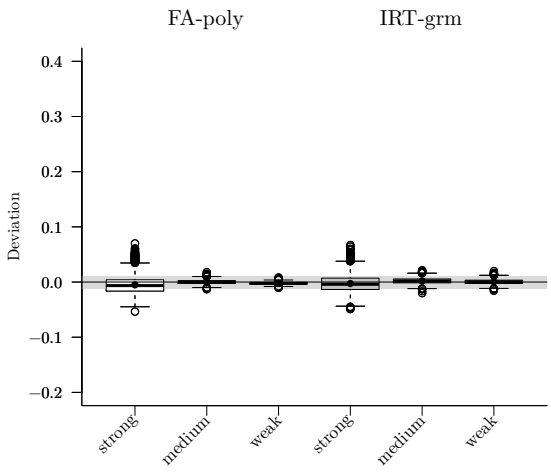



*Figure 5.19.* $\hat{se}(\hat{\alpha}_i) - sd(\hat{\alpha}_i)$ for FA-poly and IRT-grm. The grey area represents an approximation of the margin of deviation around the true value considered acceptable. $n = 600$; $R = 4000$.

**Step-Difficulty Parameters** In the MANOVA applied to the PB of step-difficulty parameter estimators no significant effects were found larger than 0.01, so no MANOVA table is presented here. Unsurprisingly, these results resemble those of the threshold parameter estimates (cf. Table 5.4).

For the medium sample size, the estimates are depicted graphically in Figure 5.20. The true values are given as thin lines crossing each box. The small sample size results are not shown, as the variance is even larger there (outliers of +40 and −40 for the outer step-difficulties of the weak loading items), making that graph unreadable.

Despite the lack of large significant effects, we can observe a pattern in Figure 5.20: IRT-grm step-difficulty estimators of an absolute value smaller than 2, approximately, are biased towards zero, whereas otherwise they are biased away from zero. The bias is larger for the more extreme values, which is in accordance with Expectations 4c and 4d.

FA-poly estimators are all biased away from zero, i.e., positively signed step-difficulty parameters are overestimated and negatively signed step-difficulty parameters are underestimated. Since step-difficulty parameters are larger for smaller item loadings, these effects are also related to item strength, in the sense that the weak-item step-difficulty values are more dispersed over the LV scale. A rather large bias of 0.55 and 0.80 in absolute value is present for the outer thresholds of the weak item for

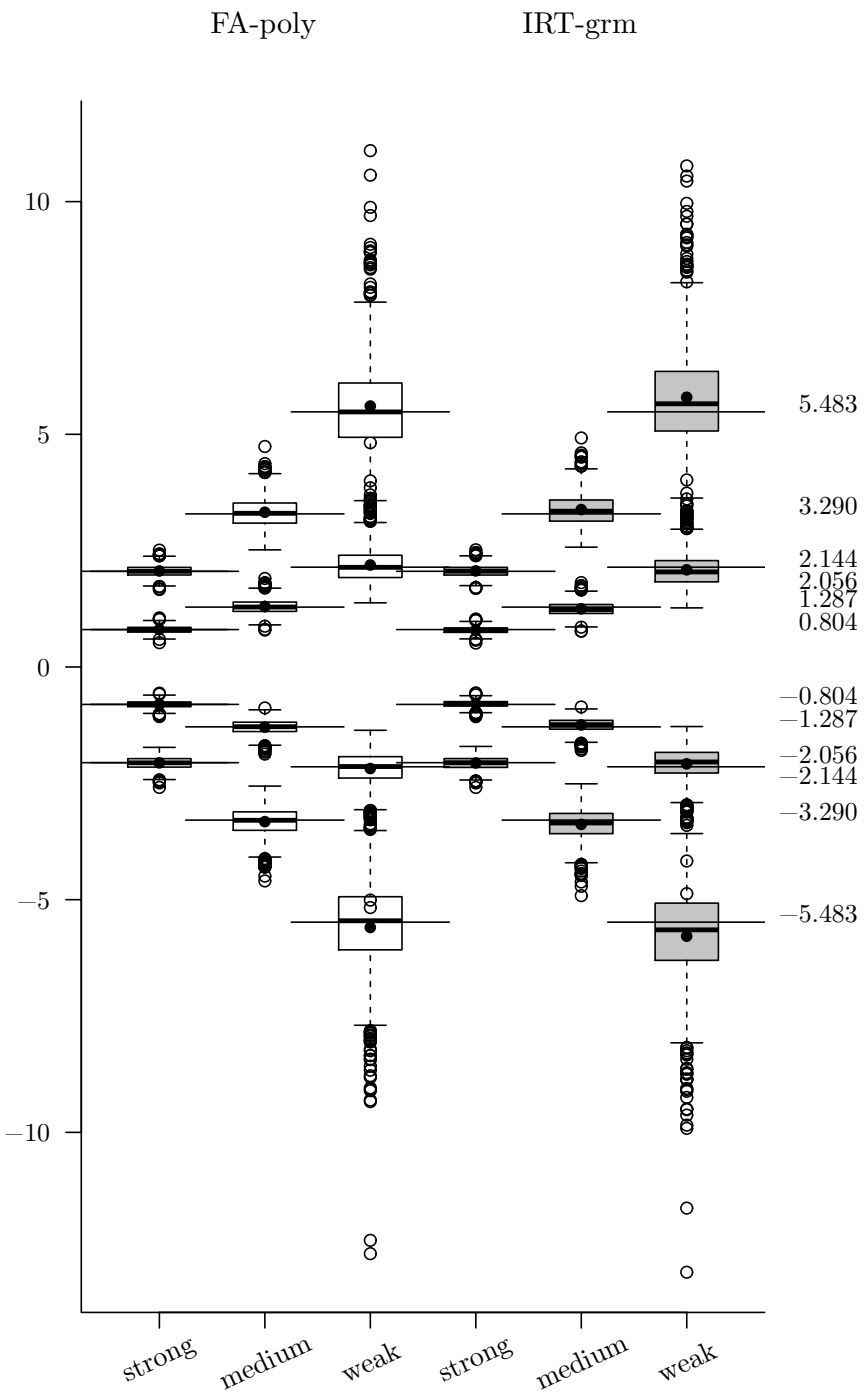



*Figure 5.20.* Parameter estimates $\hat{\beta}_{ic}$ for strong, medium, and weak items, as estimated by FA-poly and IRT-grm. The horizontal lines represent the true values, identified in the right margin. $n = 600$; $R = 1000$.

FA-poly and IRT-grm, respectively. As this concerns the plain bias, these values are on the LV scale (cf. Figure 4.1).

In addition, we observe a larger variance of estimates for more extreme step-difficulty parameters. When comparing the estimates of the outer thresholds of the strong item with the inner thresholds of the weak item, we observe that, even though their values on the latent scale are about the same (e.g., $\beta_{1.4} = 2.056$ and $\beta_{9.3} = 2.144$), the inner threshold estimates for the weak item demonstrate more variance.

### Coverage Rates

As the coverage results much resemble the parameter estimation results, they are only dicussed very briefly. Coverage rates of the 95%-confidence interval of loading and threshold parameter estimators for all parametric models in the basic data configurations are given in Table D.14 of Appendix D.4.

For FA-poly and IRT-grm, most coverage rates are about 0.95, which is as it should be. For FA-poly, the worst — though still acceptable — average coverage rates are for the medium and weak loading parameter when the sample size is small (deviation of $-0.028$ and $-0.023$, respectively). For IRT-grm, the inner thresholds of items loading medium or weak on the LV have unacceptable coverage rates in case of the medium sample size (deviations ranging from $-0.054$ to $-0.140$).

For FA-lin, the majority of coverage rates deviate unacceptably from the expected 0.95. The worst result is found for the strong loading in case of the medium sample size, with a deviation of $-0.597$.

### Summary of Parameter and Standard Error Results

*Loading* parameter estimators are most severely and consistently biased for FA-lin, with an RB of about $-6\%$ (consistent with Expectations 1b and 2a). FA-poly and IRT-grm loading parameter estimators are both unbiased (in accordance with Expectations 3a and 4a). However, IRT-grm estimators show an increasingly negative bias for decreasing parameter values, whereas FA-poly estimators are unbiased over the various population values. All loading standard error estimators are considered acceptable by our 10% RB criterion. Comparing the models, FA-poly standard error estimators demonstrate the largest bias, with an RB of as much as $-7.5\%$ for a weak item (consistent with Expectation 3l). As we observe no standard error bias for FA-lin, our standard error estimation results are better than those reported by Rhemtulla et al. (2012) for $n = 150$, on which we based our tentative Expectation 2e of standard error bias for the small sample size, and more in line with their $n = 350$ results. It seems that a sample size of $n = 200$ is large enough for unbiased FA-lin standard error estimators.

For FA-poly, *threshold* parameter estimators are unbiased (in accordance with Expectation 3b), whereas for IRT-grm the PB of the inner thresholds of medium and weak items is relatively large. The latter was unexpected based on previous step-difficulty estimation results (Expectations 4c and 4d). However, step-difficulty results are con-

sistent with these expectations. As step-difficulty parameters are dependent on both loading and threshold parameters, findings regarding step-difficulties do not generally hold equally for thresholds. Standard error estimators of threshold parameters are unbiased for both models and in each condition.

FA-poly and IRT-grm *discrimination* parameter estimators are generally unbiased. For FA-poly this is consistent with the results for item loadings and thus in accordance with Expectation 3a. For IRT-grm we found an increasing, though still acceptable, bias with a decreasing discrimination parameter, which is in line with our tentative Expectation 4b. Therefore, we did not find an increasing bias for increasing item discrimination, as reported by Boulet (1996) and Finger (2001). Discrimination standard error estimators are considered acceptable for both models in each condition. As for the loadings, FA-poly discrimination standard errors demonstrate the most bias, with an RB of at most $-7.7\%$ for a medium loading item.

The bias of *step-difficulty* parameter estimators is more severe for IRT-grm than for FA-poly, and bias increases with increasing sample size and decreasing item loadings, which is in accordance with Expectations 4c and 4d.

Generally, the medium sample size ($n = 600$) results in less variability of parameter and standard error estimates than the small sample size ($n = 200$) for all parametric models (consistent with Expectation 1i). The bias of estimators is not affected by sample sizes, except for the step-difficulty parameters, where the bias is smaller for the medium sample size.

The coverage of the FA-lin loading estimator confidence interval for a strong item is unacceptably low, especially for the medium sample size. IRT-grm coverage rates are unacceptable for the inner thresholds of items loading medium or low on the LV. FA-poly coverage rates are all acceptable.

In conditions of normal items loading on a normal LV, FA-poly seems to perform best in terms of parameter and standard error estimation, closely followed by IRT-grm. FA-lin is clearly outperformed.

### 5.2.4 Fit Indices

*Table 5.6.* ANOVA results: $\eta^2$ per effect for RMSEA and SRMR fit statistics. $N = 12000$.

| Effect | Levels | $\eta^2$ for RMSEA | $\eta^2$ for SRMR |
|---|---|---|---|
| Model (m) | 3 | | 0.155 |
| Scale strength (s) | 2 | | 0.416 |
| Sample size (n) | 2 | 0.064 | 0.330 |
| s × n | 4 | | 0.018 |

*Note.* Listed effects are statistically significant at $\alpha = 0.01$ and are sized $\eta^2 > 0.01$.

*Table 5.7.* Fit statistics for Cells nNS2, nNS6, nNM2, and nNM6 as estimated by the parametric models, averaged over replications. $R = 1000$.

| Cell | Fit Statistic | FA-lin $df = 54$ | FA-poly $df = 18$ | IRT-grm $df = 18$ |
|---|---|---|---|---|
| nNS2 | $\chi^2_{YB}$ | 57.229 | 18.469 | 18.576 |
| | RMSEA | 0.016 | 0.017 | 0.018 |
| | SRMR | 0.026 | 0.026 | 0.032 |
| nNS6 | $\chi^2_{YB}$ | 55.790 | 18.361 | 18.304 |
| | RMSEA | 0.008 | 0.010 | 0.010 |
| | SRMR | 0.015 | 0.015 | 0.023 |
| nNM2 | $\chi^2_{YB}$ | 57.137 | 18.727 | 18.971 |
| | RMSEA | 0.017 | 0.018 | 0.019 |
| | SRMR | 0.043 | 0.043 | 0.052 |
| nNM6 | $\chi^2_{YB}$ | 55.822 | 18.476 | 18.480 |
| | RMSEA | 0.008 | 0.010 | 0.010 |
| | SRMR | 0.025 | 0.025 | 0.039 |

To investigate the effect of the explanatory variables on model fit, we applied ANOVAs to the RMSEA based on the $\chi^2_{YB}$ and to the SRMR, the results of which are presented in Table 5.6. The RMSEA results are listed in the second column of Table 5.6, where we can observe an effect of sample size ($\eta^2 = 0.064$).

In Figure 5.21 the distribution of the RMSEA is compared to its expected counterpart for each parametric model and each cell of the design. In the inset of the subfigures, the degrees of freedom, the NCP (see Equation 4.24), the mean, and the variance of the distribution of estimates is given. The theoretical values were generated using the estimated NCP. In Appendix D.6, the observed and expected $\chi^2_{YB}$ values are presented in a similar figure. Overall, the estimated distributions are similar to the expected distributions. FA-lin stands out slightly — though the differences are small — by showing the largest deviation of the theoretical values in the tails of the distribution when sample size is small.

In the last column of Table 5.6, the results of the ANOVA applied to the SRMR estimates are presented. The main effect of scale strength ($\eta^2 = 0.416$) is the largest, followed by sample size ($\eta^2 = 0.330$), model ($\eta^2 = 0.155$), and a smaller interaction effect of scale strength and sample size ($\eta^2 = 0.018$). From Table 5.7, providing fit statistics for each design cell averaged over replications, we see that the SRMR is larger for the mixed than for the strong scale, and larger for the small than for the medium sample size. Moreover, values are slightly larger for IRT-grm than for the FA models. This indicates a better model-data fit for the strong scale, the medium sample size, and the FA models.

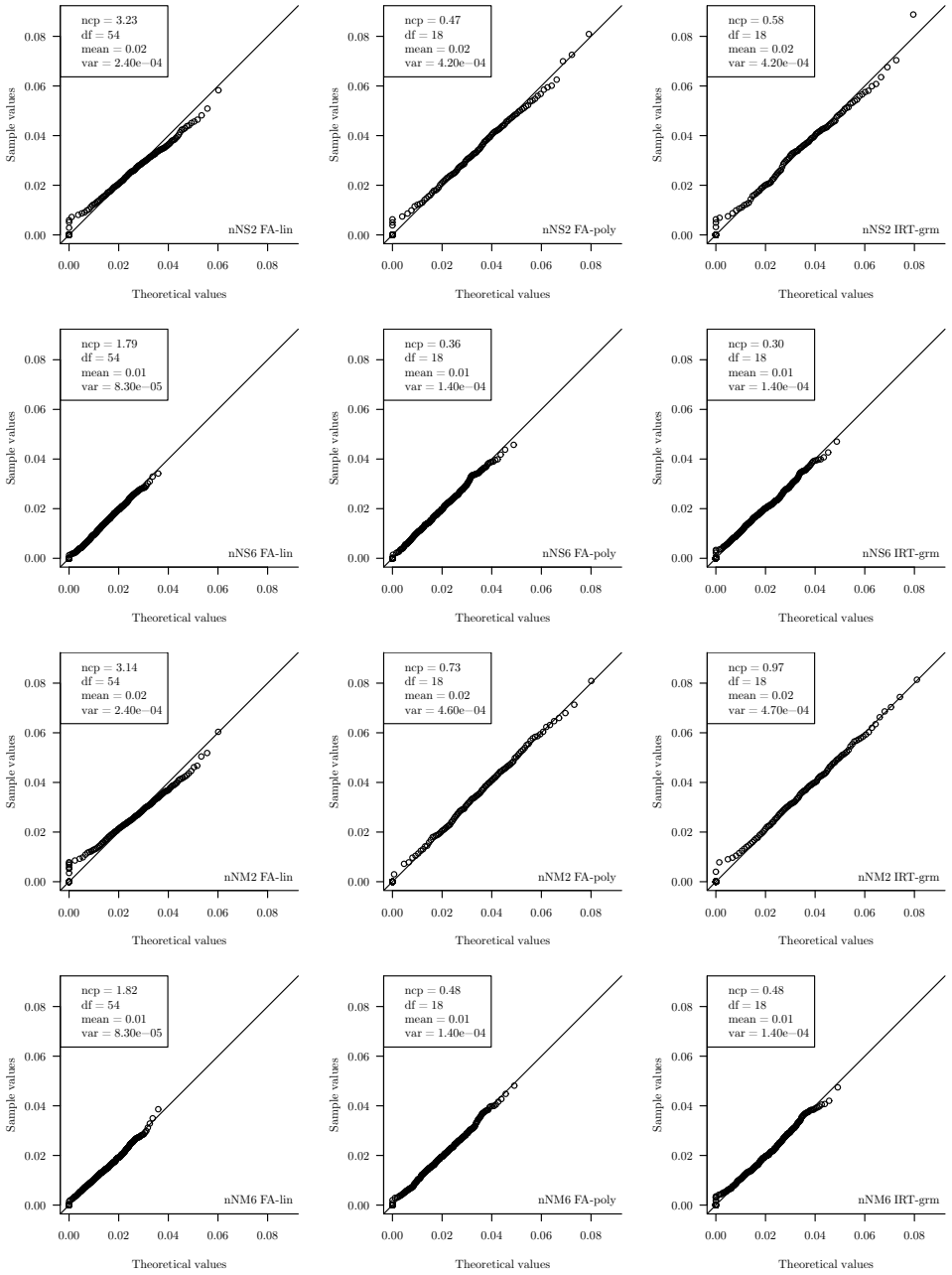



*Figure 5.21.* Q-Q plots for RMSEA fit statistic for Cells nNS2, nNS6, nNM2, and nNM6 and each model. $R = 1000$. The diagonal line depicts a perfect association between the empirical and theoretical distribution, the latter being a noncentral $\chi^2$ distribution using the mean empirical noncentrality parameter (NCP) over $R$ replications.

In addition to the SRMR, Table 5.7 provides the $\chi^2_{YB}$ and the RMSEA. As the degrees of freedom are the expected values of the $\chi^2_{YB}$ distribution, we can conclude that both FA-poly and IRT-grm $\chi^2_{YB}$ estimators are unbiased, which is in accordance with Expectations 3q and 4q, respectively. FA-lin $\chi^2_{YB}$ values are overestimated slightly but acceptable and more so for the small than the medium sample size, which supports our tentative Expectation 2g, based on the good performance of the FA-lin $\chi^2_{YB}$ statistic in case of continuous items reported by Maydeu-Olivares et al. (2011).

RMSEA values are, generally, smaller for the medium than for the small sample size (respective means of 0.01 and 0.02), indicating a better fit for the medium sample size. Both by the RMSEA criterion of $< 0.06$ and the SRMR criterion of $< 0.08$, all models fit the data well in each condition, which is consistent with Expectations 2f, 3s, and 4s.

### 5.2.5 Nonparametric IRT-mok

The results of applying the nonparametric IRT-mok model are discussed next. We focus on the estimation of the scalability coefficient Loevinger's $H$ and corresponding standard errors, on item and scale level. As the data were configured according to the FA-poly/IRT-grm model, we first present the population $H$ values, which were derived from the loading and threshold configuration.

In Table 5.8 the true Loevinger's $H_i$ and $H_{scale}$ values for the design cells are listed. The strong scale of Cells nNS2/6 has an $H_{scale}$ value of 0.571, and the mixed-strength scale of Cells nNM2/6 has an $H_{scale}$ of 0.250.

Comparing the $H$ values to the loading parameter values, we gather that a scale that is acceptable according to a common FA criterion, i.e., all $\lambda_i \geq 0.3$, is judged unacceptable by the standard IRT-mok criterion, i.e., all $H_i \geq 0.3$, and thus $H_{scale} \geq$

*Table 5.8.* IRT-mok $H_i$ and $H_{scale}$ true values for each cell of the design.

| | nNS2<br>nNS6 | nNM2<br>nNM6 |
|---|---|---|
| $H_1$ | 0.571 | 0.363 |
| $H_2$ | 0.571 | 0.363 |
| $H_3$ | 0.571 | 0.363 |
| $H_4$ | 0.571 | 0.363 |
| $H_5$ | 0.571 | 0.239 |
| $H_6$ | 0.571 | 0.239 |
| $H_7$ | 0.571 | 0.239 |
| $H_8$ | 0.571 | 0.239 |
| $H_9$ | 0.571 | 0.148 |
| $H_{10}$ | 0.571 | 0.148 |
| $H_{11}$ | 0.571 | 0.148 |
| $H_{12}$ | 0.571 | 0.148 |
| $H_{scale}$ | 0.571 | 0.250 |

*Table 5.9.* MANOVA results: $\eta_p^2$ per effect for RB of $H_i$ parameters and of corresponding standard errors. $N = 4000$.

| Effect | Levels | Parameters $\eta_p^2$ | Standard errors $\eta_p^2$ |
|---|---|---|---|
| Scale strength (s) | 2 | | 0.105 |
| Sample size (n) | 2 | 0.017 | |
| s × n | 4 | | 0.029 |
| Item group (ig) | 3 | | |
| s × ig | 6 | | 0.087 |
| n × ig | 6 | | 0.094 |
| s × n × ig | 12 | | 0.046 |

*Note.* Listed effects are statistically significant at $\alpha = 0.01$ and are sized $\eta_p^2 > 0.01$.

0.3. The scalability of an item $H_i$ is directly related to the other items in the scale, causing an item with a strong relation to the LV to obtain a relatively small $H_i$ as a result of the presence of other items in the scale that have a weak relation to the LV. This is an important difference compared to the loading/discrimination parameters of the parametric models, as the latter are not characterized by such interdependencies. We therefore emphasize that, due to the weak relation between the weak items and the LV, the condition of the mixed scale is certainly not favorable in an IRT-mok analysis.

**Parameters** The results of the MANOVA applied to the RB-constituents of the $H_i$ parameters and their standard error estimators are presented in Table 5.9. For the parameters, we see a small effect of sample size ($\eta_p^2 = 0.017$).

In Figure 5.22 the deviation of the $H_i$ estimates from the true $H_i$ values is plotted per cell of the design per applicable item type, resulting in one plotted item for the strong scales and three plotted items for the mixed-scale cells. Note that the strong item of Cells nNS2/6 has a larger true $H_i$ value than a strong item in Cells nNM2/6, because item strength is defined from the loading specification of the entire scale (see also Table 5.8).

Figure 5.23 presents the deviation of $H_{scale}$ estimates in each cell of the design. For the strong scale, the RB of $H_i$ and $H_{scale}$ equals 4.8% for the small sample size, and 2.9% for the medium sample size (see Tables E.65 and E.66). For the mixed scale, the RB is 5.2% for the small sample size, and 3.3% for the medium sample size (see Tables D.17 and D.18).

In general, the $H$ estimates are too large compared to their population values. The deviation decreases with increasing sample size. In fact, when we ran a simulation with an extremely large sample size ($n = 500000$) the $H$ estimates converged to the population values. The positive bias of $H$ values was not expected, and might call for some caution in interpreting sample $H$ values in scale analysis.

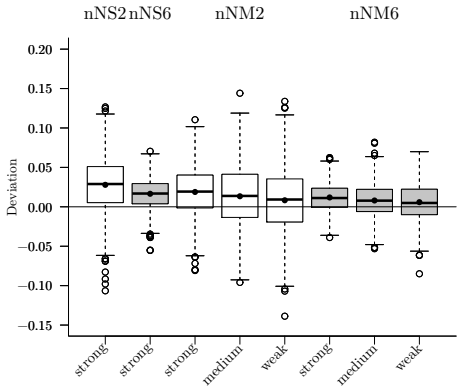



*Figure 5.22.* $\hat{H}_i - H_i$. $n = 200$ for Cells nNS2 and nNM2; and $n = 600$ for Cells nNS6 and nNM6. $R = 1000$ for the boxplots of each cell.

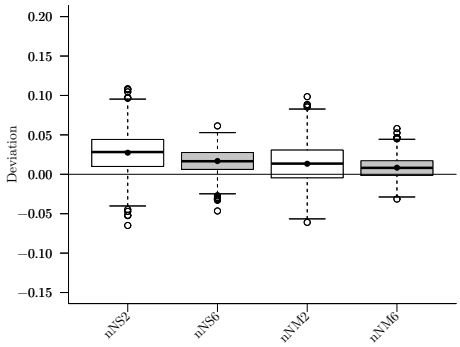



*Figure 5.23.* $\hat{H}_{scale} - H_{scale}$. $H_{scale} = 0.571$ for Cells nNS2 and nNS6; and $H_{scale} = 0.25$ for Cells nNM2 and nNM6. $n = 200$ for Cells nNS2 and nNM2; and $n = 600$ for Cells nNS6 and nNM6. $R = 1000$.

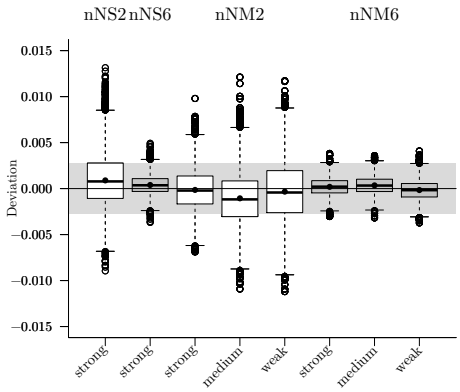



*Figure 5.24.* $\hat{se}(\hat{H}_i) - sd(\hat{H}_i)$. $n = 200$ for Cells nNS2 and nNM2; and $n = 600$ for Cells nNS6 and nNM6. $R = 12000$ for the boxplots of Cells nNS2 and nNS6; and $R = 4000$ for the boxplots of Cells nNM2 and nNM6. The grey area represents an approximation of the margin of deviation around the true value considered acceptable.

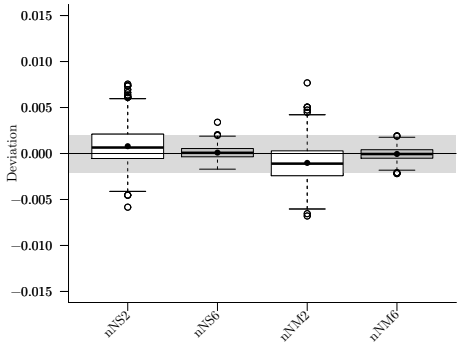



*Figure 5.25.* $\hat{se}(\hat{H}_{scale}) - sd(\hat{H}_{scale})$. $n = 200$ for Cells nNS2 and nNM2; and $n = 600$ for Cells nNS6 and nNM6. $R = 1000$. The grey area represents an approximation of the margin of deviation around the true value considered acceptable.

**Standard Errors and Coverage** Standard error results are reported in the last column of Table 5.9. The largest effect is of scale strength ($\eta_p^2 = 0.105$). In addition,

the interaction effect of scale strength, sample size, and item group ($\eta_p^2 = 0.046$) is worth mentioning, being the most complicated significant interaction. This implies that when the sample size is small, the standard errors of $H_i$ are overestimated for items of the strong scale and underestimated for the medium items of the mixed scale, as can be observed in Figure 5.24, where the deviation of the standard error estimates from the parameter standard deviation is graphically depicted for $H_i$. An $H$ value of a mixed-scale item is thus overestimated with an overestimated precision. In an applied setting, this results in a too optimistic impression of item scalability with possible implications for the retention of items in a scale.

Standard error estimates of $H_{scale}$ are presented in Figure 5.23. All standard error RB values are smaller than 5% and are thus well within our 10% bound of acceptable deviation.

The average coverage rate of $H_i$ is 0.86 for items in the strong scale, regardless of the sample size, and therefore considered too small. For the mixed scale, coverage rates are acceptable as they vary between 0.90 and 0.94.

**Summary** $H$ values are consistently overestimated. This bias decreases with increasing sample size. Standard errors are all considered acceptable. As Monte Carlo research including IRT-mok is lacking, these results are unprecedented.

### 5.2.6 Latent Variable Score Estimates

In Table 5.10 results are presented of the ANOVA applied to Kendall's $\tau_a$ between the true and estimated LV scores. LV results of all four models are included in this analysis, as is apparent from the number of levels for the explanatory variable model. We observe a large main effect of scale strength ($\eta^2 = 0.864$), and a small interaction effect of model and scale strength ($\eta^2 = 0.052$).

The associations between the LV true scores and their estimates are illustrated by the scatterplots of Figure 5.26 for all models in all design cells. The deviation scores $\theta_{sr} - \bar{\theta}_r$ and $\hat{\theta}_{sr} - \bar{\hat{\theta}}_r$ of the true and estimated scores, respectively, are depicted to focus on the within-sample variations and cancel out the average variation between

*Table 5.10.* ANOVA results: $\eta^2$ per effect for Kendall's $\tau_a$ between true and estimated LV scores. $N = 16000$.

| Effect | Levels | $\eta^2$ for Kendall's $\tau_a$ |
|---|---|---|
| Model (m) | 4 | |
| Scale strength (s) | 2 | 0.864 |
| Sample size (n) | 2 | |
| m × s | 8 | 0.052 |

*Note.* Listed effects are statistically significant at $\alpha = 0.01$ and are sized $\eta^2 > 0.01$.

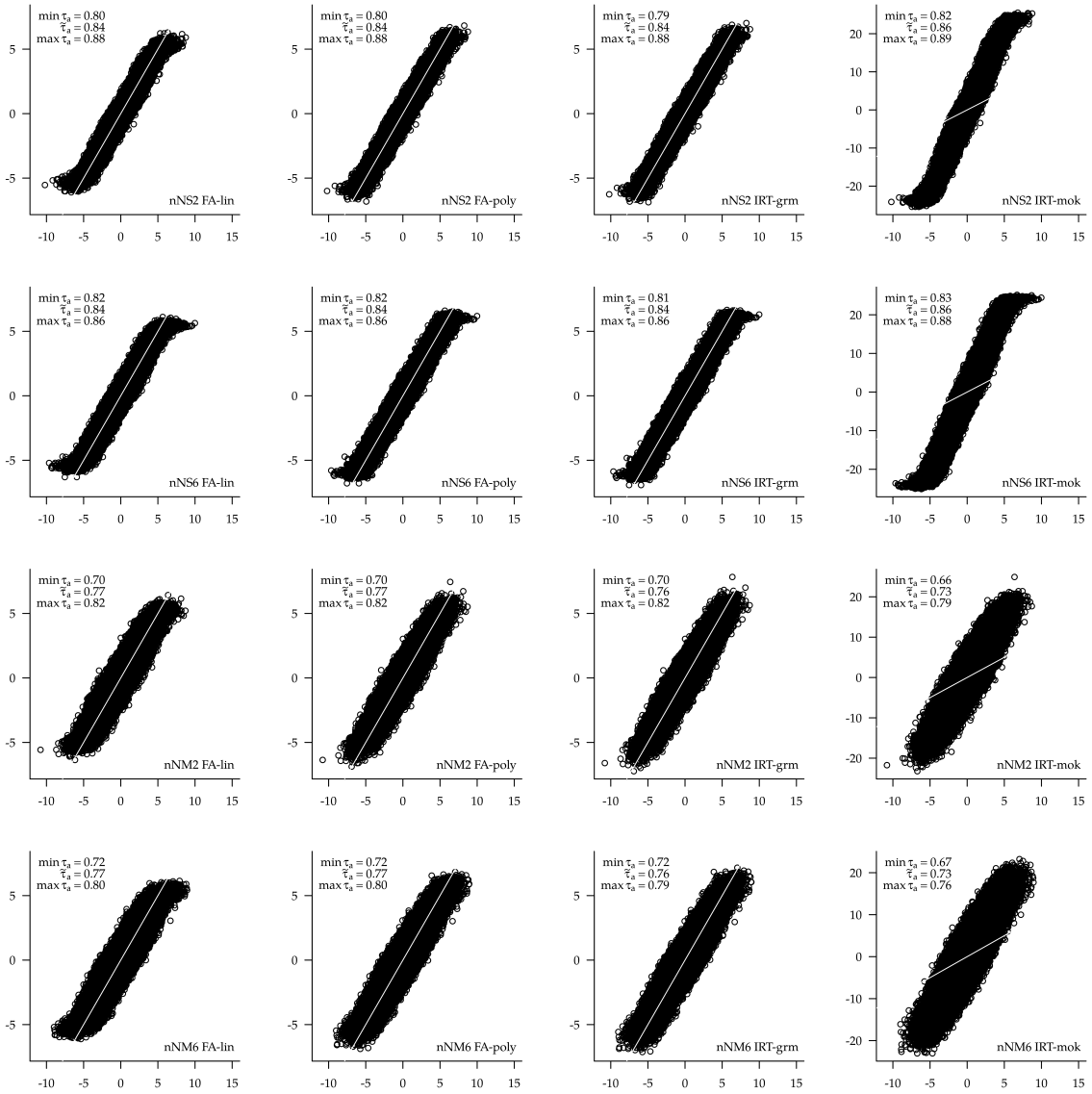



*Figure 5.26.* Scatterplots of LV population $\theta_{sr} - \bar{\theta}_r$ ($x$-axis) and estimated $\hat{\theta}_{sr} - \bar{\hat{\theta}}_r$ ($y$-axis) deviation scores for FA-lin, FA-poly, IRT-grm, and IRT-mok in Cells nNS2, nNS6, nNM2, and nNM6 for each replication. The minimum, median, and maximum Kendall's $\tau_a$ over replications are given in the inset of each plot as an indication of association. $n = 200$ for Cells nNS2 and nNM2; and $n = 600$ for Cells nNS6 and nNM6; $R = 1000$.

replications. The scatterplots of every single replication are overlaid in each of the subfigures. The white lines depict the identity association. Because IRT-mok LV scores are not on a latent scale, but on the integer scale of sum scores, the white line is not approximated well by the scatterplots, whereas this is the case for the parametric models. The minimum, median, and maximum Kendall's $\tau_a$ are given in the inset of each subplot.

LV score estimation seems to be similar for all estimation models, which differs from Expectation 1h, where we expected better performance of FA-poly and IRT-grm than of FA-lin and IRT-mok, especially in the tails of the distribution based on an empirical study by Dumenci and Achenbach (2008). Judging from the scatterplots, estimation is worst in the tails of the distributions for all models, which is in accordance with Expectations 2k and 4t.

For the strong scale, the association between the true and estimated LV scores is greater (median Kendall's $\tau_a = 0.85$) than for the mixed scale (median Kendall's $\tau_a = 0.77$). When we ran simulations with all medium item loadings, we found a further decline in Kendall's $\tau_a$ to 0.70 for each model.

For IRT-mok, LV score estimates are unweighted sum scores, as is apparent from the vertical axis starting ranging from $-20$ to 20. Furthermore, the strong-versus-mixed scale difference is present even more strongly for this model, with median Kendall's $\tau_a$ equal to 0.86 and 0.73, respectively. Note that, in case of the strong scale, the association between estimated and true LV scores is stronger for IRT-mok than the other models. For the mixed scale, the benefit of weighting items by their estimated loading becomes apparent: IRT-mok LV scores are based on all items weighted equally, whereas the other models attribute more importance to items that are related to the LV in a stronger sense, thus composing a more accurate LV score estimate.

In summary, scales consisting of items that are all strongly related to the LV produce better LV estimates than either mixed or medium scales.

## 5.3 Discussion

In this chapter, we presented the results of applying the four scaling models to four basic data configurations. We compared FA of the sample covariance matrix (FA-lin), FA of the estimated polychoric correlation matrix (FA-poly), the graded response IRT model (IRT-grm), and the nonparametric Mokken IRT model (IRT-mok) on a number of performance criteria regarding estimators of parameters, corresponding standard errors, model fit, and latent variable (LV) scores. Data were generated under the FA-poly model, with both item and latent variable (LV) distributions being normal. The four data configurations differed in (a) the strength of the scale: all items loading $\lambda = 0.80$ on the LV (strong scale) versus items loading either $\lambda = 0.80$, $\lambda = 0.50$, or $\lambda = 0.30$ on the LV (mixed scale), and (b) the sample size: $n = 200$ (small) versus $n = 600$ (medium).

As was expected, only small differences between the models were found, since distributional assumptions were met in case of FA-poly and IRT-grm, and only mildly

*Table 5.11.* Summary of results from normal data configurations, referring to the expectations presented in Section 4.3 (p. 91ff.). Results deviating from the expectations are printed larger and in boldface.

| | LV | Normal distribution | | | |
|---|---|---|---|---|---|
| | | $\hat{\omega}$ | $\hat{se}(\hat{\omega})$ | Model | LV score |
| Item | Model | Bias | Bias | Fit | Bias |
| Normal | FA-lin | − | **√**[a] | √ | **√/−**[b] |
| | | 2a | **2e** | 2f/2g | **2k** |
| | FA-poly | √ | √ | √ | √ |
| | | 3a/3b | 3l/3n | 3q/3s | 3t |
| | IRT-grm | √ | √ | √ | √ |
| | | 4a/4b/4c/4d | 4l/4n | 4q/4s | 4t |

*Note.* √ indicates good performance; + and − indicate positive and negative bias, respectively.

[a] FA-lin loading standard error estimators were unbiased, although we expected them to be negatively biased. [b] FA-lin LV score estimates were expected to deviate from their true counterparts, but a strong association was found.

violated in case of FA-lin, supposing continuous item variables that were in fact ordered categorical. The results offer a frame of reference useful for the interpretation and evaluation of the results of our additional data configurations with violations of distributional assumptions, covered in Chapter 6.

In Table 5.11 a summary of the results regarding the parametric models is given by referring to our expectations presented in Chapter 4 and summarized in Table 4.2 (p. 98). From the table it can be observed that most of our expectations with regard to the normal data configurations were supported. Two exceptions are the expectations regarding FA-lin standard error and LV score estimators, which were found to be more accurate than expected. Furthermore, it is clear that both FA-poly and IRT-grm perform well with respect to all the performance variables. In the following, we will elaborate on these findings.

### Parameters and Standard Errors

Loading parameter estimators were unbiased for FA-poly and IRT-grm (consistent with Expectations 3a and 4a). Although within the acceptable range of 5% deviation, we found more bias for IRT-grm than for FA-poly, and more so for the weaker items. FA-lin consistently produced biased loading parameters, with a relative bias of −6%, which is in accordance with Expectation 2a.

All loading standard error estimators showed acceptable performance, with FA-poly estimators displaying the most bias, of maximally −7.5%, which is concordant with Expectations 3l and 4l. As a consequence, significance tests of a parameter estimate are expected to be liberal.

Because the literature was inconclusive with respect to the quality of FA-lin standard error estimators when applied to categorical data, we tentatively expected some bias for our small sample size (Expectation 2e). Nevertheless, we found no standard error bias, hence our results were better than reported in previous research by Rhemtulla et al. (2012), who applied robust corrections for nonnormality to the standard error estimates, and found underestimated loading standard errors for sample sizes of 100 and 150. For the larger sample sizes of 350/600 included in their study, however, Rhemtulla et al., did not find substantial standard error bias. This is in line with DiStefano (2002), who reported no standard error bias for sample sizes of 350 and 700. Apparently, a sample size of 200 is large enough for unbiased FA-lin standard error estimators. However, one should note the limited relevance of accurate standard error estimators in case of biased parameter estimators, which applies to FA-lin.

Discrimination parameters were unbiased for both FA-poly and IRT-grm. For FA-poly, this is consistent with the results for item loadings and thus in accordance with Expectation 3a. For IRT-grm we found an increasing RB with a decreasing discrimination parameter value — though smaller than 5% — which was expected based on Forero and Maydeu-Olivares (2009), but contrary to findings of Boulet (1996) and Finger (2001), who reported an increase in bias with increasing item discrimination (Expectation 4b). As the the latter two studies concerned dichotomous items, we also ran simulations with dichotomous items. In these additional runs, however, we found results similar to our polytomous item results. Perhaps the differences in results are caused by the various software used in the studies, as Boulet (1996) used TESTFACT, Finger (2001) used his own software routines, and Forero and Maydeu-Olivares (2009) used MPLUS, as we did.

Standard error results of the discrimination parameters highly resembled their loading counterparts, with no substantial bias, largest for FA-poly (at most −7.7%).

FA-poly threshold parameters were unbiased, which is in accordance with Expectation 3b. IRT-grm threshold parameter estimators were more biased for the inner than for the outer thresholds for the medium and weak items, which was unexpected given previous step-difficulty estimation results (Expectations 4c and 4d), but this bias was small. The IRT-grm results for step-difficulty parameter estimators, however, were congruent with these expectations, with more severe bias for increasing sample size and decreasing item loadings. FA-poly step-difficulty parameter estimators were all biased opposite from zero. For FA-poly and IRT-grm a rather large bias of 0.55 and 0.80 in absolute value, respectively, was found for the outer thresholds of weak items. Step-difficulty parameters were the only item parameters that demonstrated a substantial decrease in bias with an increasing sample size.

### Model Fit

Model fit, as indicated by the $\chi^2_{YB}$, RMSEA, and SRMR, was good for each of the estimation models in every condition, supporting Expectations 2f, 2g, 3q, 3s, 4q, and 4s, with small effects of a better fit for a larger sample size and a stronger scale, and a better fit for the FA models compared to IRT-grm, as indicated by the SRMR.

### Nonparametric IRT-mok Model

We found the scalability coefficient $H$ to be consistently positively biased. This overestimation decreased with increasing sample size, approximating zero for an approximately infinite sample size. This is remarkable, as nonparametric modeling is usually employed when the available sample size is small, and advertised for such cases. The reported bias of 5% is not extreme, but does call for some caution in interpreting the scalability coefficient. The observed variance of the $H$ estimates supports this call for caution, for example, when applying cut-off criteria for retaining items in a scale.

Standard errors for $H$, recently made available by Kuijpers, Van der Ark, and Croon (2013), were found to be unbiased. Consequently, they are a very useful aid in the interpretation of $H$ coefficients in scale analysis, taking into account the variability of the coefficient.

### LV Scores

For all scaling models alike, the LV score estimates were strongly associated (Kendall's $\tau_a \approx 0.8$) with their true counterparts, which is consistent with Expectations 3t and 4t for FA-poly and IRT-mok, and better than expected for FA-lin in Expectation 2k.

Only in the tails of the distribution did the models show some deviations, but these results were similar for all models and do not support Expectation 1h, based on Dumenci and Achenbach (2008) who found better results for FA-poly and IRT-grm than for FA-lin and unweighted sum scores in their empirical study. Perhaps our findings are different because the item distributions in that study were all moderately to highly skewed. In the next chapter we shall examine LV score estimation in case of item and LV skewness.

The LV estimates for the mixed scale were worse for IRT-mok as compared to the other models, which underlines the benefits of weighting items by their estimated loadings. Whether this also holds when the assumptions of the parametric models are violated will become clear in Chapter 6.

### Sample Size

Results were, generally, better for the medium ($n = 600$) than for the small ($n = 200$) sample size, which is consistent with Expectation 1i. Parameter and standard error bias did not substantially improve going from the small to medium sample size, except for the step-difficulty and $H_i$ parameters. Precision of all parameter and standard error estimators much improved with increasing sample size.

Whether the small sample size of $n = 200$ is generally large enough depends on the objectives of the researcher. Parameter estimates were quite variable for $n = 200$, gathering from the relatively large variance of parameter estimators.

## 5.4 Recommendations

Based on the results of the first part of our Monte Carlo design, concerning normal LV and item distributions, we provide the following recommendations to applied researchers performing a scale analysis on approximately normal ordered categorical data in case the LV can be assumed to be approximately normal.

- When selecting an estimation model for scale analysis regarding ordered categorical items, FA-poly-WLSMV or IRT-grm-MLR are to be preferred over FA-lin-ML.
- Loading and threshold parameter estimators of FA-poly-WLSMV and IRT-grm-MLR are considered useful and informative. Loading standard errors tend to be underestimated by FA-poly-WLSMV, calling for some caution in interpreting tests of significance, as they could be liberal. Items ought not be judged using significance testing only; their content should always be taken into account.
- The $\chi^2_{YB}$ fit statistic, the RMSEA based on it, and the SRMR can be used for assessing model fit for FA-poly-WLSMV and IRT-grm-MLR, as it produced good results. It should be mentioned, however, that behavior of these statistics in case of model *mis*fit, i.e., items do not (or only minimally) load on an LV, was not investigated here.
- The nonparametric IRT-mok produces useful results, although $H$ values are slightly overestimated for limited sample sizes. Standard errors are unbiased and can be a useful aid in interpreting the parameter estimates.
- For the strong scale, LV scores were estimated equally well by all models included in the study. The ordering of respondents based on their LV scores was more accurate for the strong than for the mixed-item scale, especially for the IRT-mok LV scores, which do not take into account the variations between items with regard to the item-LV association. This affirms the advantages of strong items in a scale and of weighting items by their estimated loading parameter.
- If precision of parameter estimates is not of great importance, a small sample size will do. However, in case one is especially interested in the strength of the item-LV relations, a larger sample size is highly recommended.

In the next chapter we will investigate the differences between the models under investigation in conditions of a nonnormal LV and item distributions.

# Chapter 6

# Simulation Study: Violations of Assumptions

In the previous chapter, we presented results of four scaling models, factor analysis of the sample covariance matrix (FA-lin), factor analysis of the estimated polychoric correlation matrix (FA-poly), the graded response item response theory model (IRT-grm), and the nonparametric Mokken item response theory model (IRT-mok), applied to categorical data configurations with approximately normal item distributions and a normal latent variable (LV) distribution.

Having established the models' performance under optimal conditions of normality, we now turn to configurations with violations of model assumptions, i.e., nonnormal LV and item distributions. These violations are chosen to represent empirical data commonly found in practice. Furthermore, the study design was set up taking into account results from previous simulation research, with the objective of testing the generalizability of some of these results, and systematically investigating the effects of combinations of a nonnormal LV and nonnormal item variables.

The factors in our Monte Carlo design are:

- Estimation model: FA-lin by means of maximum likelihood (FA-lin-ML), FA-poly by means of mean-and-variance adjusted weighted least squares (FA-poly-WLSMV), the graded response model by means of robust maximum likelihood (IRT-grm-MLR), and the nonparametric Mokken item response theory model (IRT-mok)

- LV distribution: normal and right skew-normal

- Scale shape: various combinations of normal, right-skewed, left-skewed, and bimodal item response distributions

- Sample size: small ($n = 200$), and medium ($n = 600$)

These variables are investigated in samples of data consisting of 12 five-category items loading on a single LV. In the normal data configurations discussed in the previous chapter, scale strength was also a factor in the design, varying between strong (all items load strongly [0.80] on the LV), and mixed (four items strong [0.80], four medium [0.50], and four weak [0.30]). In the present chapter, we focus on the LV and item distributions, so we keep the scale strength constant at strong. To facilitate the comparison with conditions of normality, the two design cells discussed in the previous chapter concerning strong scales are also included in the presentation and analysis of the current results.

Performance of the estimation models is evaluated using the performance variables and corresponding criteria regarding parameter estimators, corresponding standard errors, model fit, and LV scores, as explained in Chapter 4.

In the following sections, we first elaborate on the configuration of the data conditions and the setup of the meta-analyses of the results. Next, results are presented of applying the four estimation models to the samples of data. Subsequently, the results are discussed and compared to the expectations laid out in Chapter 4. And finally, recommendations are presented based on our findings.

## 6.1 Method

### 6.1.1 Data Configurations

The data configurations used to investigate the performance of the scaling models are listed in Table 6.1. As was the case for the normal cells, the cell names identify their specifications in the order: item distributions, LV distribution, scale strength, and sample size. For example, Cell *lrnNS2* consists of *l*eft-skewed, *r*ight-skewed, and *n*ormal items loading on a *N*ormal LV; all items load *S*trongly ($\lambda = 0.80$) on the LV and the sample size equals *2*00.

As mentioned earlier, scale strength is held constant to focus on the effects of deviating item and LV distributions under the preferable condition of a strong scale. Each data configuration was generated with $n = 200$ and $n = 600$. Details about the LV and items distributions included in the design were given in Chapter 4.

The LV distribution and the item distributions are manipulated independently by adjusting the item threshold values, as explained in Chapter 4. Cells rnNS2/6 and rnRS2/6 have right-skewed and normal items loading on a normal or right skew-normal LV, respectively. Cells nRS2/6 have a right skew-normal LV, but the thresholds were set to result in normal item distributions. Cells lrnNS2/6 and lrnRS2/6 also include left-skewed items. The final four cells contain bimodal and normal items, loading on either a normal LV (bnNS2/6) or a right skew-normal LV (bnRS2/6).

Combined with Cells nNS2/6, the results of which were presented in the previous chapter, the design is completely crossed, facilitating the analysis of the results by means of analysis of variance (ANOVA). Therefore, these two normal cells are also included in the presentation of results.

*Table 6.1.* Data configuration for cells of the design with violations of assumptions.

| Cell | Scale shape | LV distribution | Scale strength | Sample size |
|---|---|---|---|---|
| rnNS2 | 6 × right-skewed<br>6 × normal | normal | 12 × strong | 200 |
| rnNS6 | 6 × right-skewed<br>6 × normal | normal | 12 × strong | 600 |
| lnNS2 | 6 × left-skewed<br>6 × normal | normal | 12 × strong | 200 |
| lnNS6 | 6 × left-skewed<br>6 × normal | normal | 12 × strong | 600 |
| lrnNS2 | 4 × left-skewed<br>4 × right-skewed<br>4 × normal | normal | 12 × strong | 200 |
| lrnNS6 | 4 × left-skewed<br>4 × right-skewed<br>4 × normal | normal | 12 × strong | 600 |
| bnNS2 | 6 × bimodal<br>6 × normal | normal | 12 × strong | 200 |
| bnNS6 | 6 × bimodal<br>6 × normal | normal | 12 × strong | 600 |
| nRS2 | 12 × normal | right skew-normal | 12 × strong | 200 |
| nRS6 | 12 × normal | right skew-normal | 12 × strong | 600 |
| rnRS2 | 6 × right-skewed<br>6 × normal | right skew-normal | 12 × strong | 200 |
| rnRS6 | 6 × right-skewed<br>6 × normal | right skew-normal | 12 × strong | 600 |
| lnRS2 | 6 × left-skewed<br>6 × normal | right skew-normal | 12 × strong | 200 |
| lnRS6 | 6 × left-skewed<br>6 × normal | right skew-normal | 12 × strong | 600 |
| lrnRS2 | 4 × left-skewed<br>4 × right-skewed<br>4 × normal | right skew-normal | 12 × strong | 200 |
| lrnRS6 | 4 × left-skewed<br>4 × right-skewed<br>4 × normal | right skew-normal | 12 × strong | 600 |
| bnRS2 | 6 × bimodal<br>6 × normal | right skew-normal | 12 × strong | 200 |
| bnRS6 | 6 × bimodal<br>6 × normal | right skew-normal | 12 × strong | 600 |

### 6.1.2 ANOVA Setup

Just as we did in the previous chapter for our normal design, we performed a meta-analysis on the performance variables using either ANOVA or repeated-measures multivariate analysis of variance (MANOVA). We also included the results of the normal Cells nNS2/6 in the analyses, resulting in a full factorial design containing four explanatory variables that served as between-subject variables in the analyses: model, LV distribution, scale shape, and sample size.

For the various response variables under investigation, there were two, three, or four estimation models involved. For example, in the MANOVA of threshold parameters, only FA-poly and IRT-grm estimators are available, for loading parameters we also have FA-lin estimators, and for LV scores, we included IRT-mok sum scores in the ANOVA as well. Furthermore, there were two types of LV distributions (normal and right skew-normal), five types of scale shapes, referring to the item configurations (e.g., all normal items, or six right-skewed and six normal items, etc.), and two sample sizes (small and medium).

For the loading, discrimination, and scalability parameter estimators and corresponding standard errors, we performed a repeated-measures MANOVA on the relative bias (RB)-constituents $(\hat{\omega}_r - \omega)/\omega$ and $[\hat{se}(\hat{\omega}_r) - sd(\hat{\omega})]/sd(\hat{\omega})$, respectively, where $\hat{\omega}_r$ is the estimated parameter in repetition $r$. In addition to the between-subject variables already mentioned, we identified a within-subjects variable, item group, indicating the grouping of the item response distributions. As the item response distributions were either equal for all items, for two groups of six items, or for three groups of four items, we specified six item groups of two items.

For the threshold and step-difficulty parameter estimators, the MANOVAs were applied to the plain bias (PB)-constituents $(\hat{\omega}_r - \omega)$, because we consider deviations from the population value equally important regardless of whether in concerns extreme or middle thresholds/step-difficulties. For threshold standard error estimators the RB-constituents were taken as response variables in the MANOVA. An additional within-subjects variable, threshold type, was identified in these analyses, indicating whether the parameter is one of the outer (extreme) or one of the inner (middle) parameters.

For the model fit results, we applied an ANOVA to the standardized root mean residuals (SRMR) and to the root mean squared error of approximation (RMSEA) based on the $\chi^2_{YB}$. Kendall's $\tau_a$ of the true and estimated LV scores served as the response variable for the ANOVA of the LV estimation results.

Both the univariate and the multivariate ANOVA results are presented in terms of effect sizes. As the (M)ANOVAs are over-powered and we want to focus on the most relevant effects, we only report $\eta^2 > 0.01$ and $\eta^2_p > 0.02$ of effects that are significant at a level of $\alpha = 0.01$ for the univariate and multivariate analyses, respectively. These criteria were chosen to lead to a balanced presentation of results: not showing too many minor effects, yet still providing a sufficient amount of information. To enhance readability, the names of all main effects are listed in the tables without their effect

size, while interaction effects are omitted entirely in case of insignificant or small effects.

In the (M)ANOVA tables, the number of observations involved in the meta-analysis is indicated by $N$. To give an example, for the MANOVA on loading parameters, the crossing of design factors (3 models $\times$ 2 LV distributions $\times$ 5 scale shapes $\times$ 2 sample sizes) in combination with the number of replications ($R = 1000$) results in $N = 60000$.

## 6.2 Results

In this section, we present our findings with regard to the parameter estimates, corresponding standard errors, model fit, the nonparametric IRT-mok, and the LV score estimates. For conciseness, not all results are displayed in the text; more detailed results can be found in Appendix E.

In the following we shall frequently refer to the expectations presented in Chapter 4 (p. 91ff.). But before discussing the results they are first checked for peculiarities.

### 6.2.1 Peculiarities

In the data generating process for the cells of our design, one replication initially contained empty cells in the univariate item frequency tables, i.e., one or more response categories of one or more items were not selected by any of the simulated respondents. That replication (from Cell nRS2) was discarded and replaced with a new run.

Furthermore, the model estimation procedures converged for all samples.

### 6.2.2 Parameter and Standard Error Estimates

Parameter and standard error estimation results are discussed as we did in the previous chapter, the setup of which we briefly recollect now. To investigate both the accuracy and precision of estimators, we examine the bias of estimators and the dispersion of estimates, respectively.

In each subsection, we use three elements to present and interpret our results: a MANOVA table, PB-boxplots, and RB-values. These elements are complementary: The MANOVA tables guide us towards the most important effects; the PB-boxplots present the accuracy as well as the precision of estimators; and the RB-values provided in the text serve to focus on the comparison of the results to the criteria set in Chapter 4 (5% and 10% deviation from the population value for the parameters and standard errors, respectively).

In the presentation of our results, we distinguish between six kinds of items, referred to as item types: normal, right-skewed, left-skewed, rs-mix, ls-mix, and bimodal, where a right-skewed item is from a scale with normal and right-skewed items

only, and the rs-mix item is a right-skewed item from a scale of mixed skewnesses, i.e., with normal, right-skewed, and left-skewed items. Left-skewed and ls-mix items are defined analogously.

For the paramater estimation results, the PB-boxplots show the deviation of the parameter estimates from the true value. Each boxplot represents parameter estimates belonging to a certain item type, estimated by one of the parametric models, under the condition of a specific LV distribution. The grey area represents the $-5\%$ and $+5\%$ margin of bias considered acceptable. For example, when the true value of the loading parameters equals 0.80, the RB-criterion corresponds to a deviation ranging from $-0.04$ to $+0.04$, as marked by the grey area in the figure.

In the standard error PB-boxplots, the grey area is an *approximation* to the margin of deviation considered acceptable. Unlike for the parameters, where the population value is equal across models and design cells, for the standard errors the true value is estimated as the empirical standard deviation of the parameter estimates. Hence the acceptable margin of 10% deviation differs for each design cell and estimation model. For most parameters, the differences in empirical standard deviation between the parameter estimators of the various items were small (in the third decimal), so we took the mean empirical standard deviation over the parameters included in each figure to represent the margin of acceptable deviation, thus enhancing the interpretability of the figures. For the threshold parameters, the differences in empirical standard deviations were relatively large, so we took their maximum to approximate the margin of acceptable deviation.

When the RB of parameter or standard error estimators is presented for skewed items as a single number in the text, results of left-skewed and right-skewed items are similar and the RB given is an average value. Analogously, the RB of mixed-skewed items respresents the average over rs-mix and ls-mix items. Furthermore, the RB-values are averaged over the two sample sizes included in our design, unless they differ much between the sample sizes. The RB-values of parameter and corresponding standard error estimators reported in the following subsections are therefore averages of values listed in Tables E.1 to E.60 included in Appendix E.1. Results are listed separately there for each item type and each sample size.

### Loadings

**Parameters** We performed repeated-measures MANOVAs on the RB-constituents of the loading parameter estimators and corresponding standard errors, the results of which are presented in Table 6.2. Effect sizes are listed, significant at $\alpha = 0.01$ and exceeding a threshold of $\eta_p^2 > 0.02$.

With regard to the parameter estimators, effects are found of estimation model, scale shape, and LV distribution, and various interactions of these variables, combined with the within-subjects variable item group. The significant four-factor interaction effect of model, LV distribution, scale shape, and item group ($\eta_p^2 = 0.140$), though by far not the largest, is presumably the most meaningful effect. It indicates that skewed items are recovered poorly by FA-lin, and more so for the right skew-normal than for

the normal LV. The biased loading parameters as estimated by FA-poly in case of skewed items loading on a right skew-normal LV also contributes to that interaction effect. This becomes apparent when we turn to the graphical display of our results, used to interpret the reported effects.

In Figures 6.1 and 6.2 boxplots are shown of the deviation of the loading parameter estimates from the true value for the small and medium sample size, respectively. First, turning to the results for the normal LV conditions, it becomes apparent that both FA-poly and IRT-grm perform well, with no bias, regardless of the item type, which is consistent with our Expectation 3c for FA-poly as far as the skewed items are concerned. For the bimodal items, our results are unprecedented, as such items were not included in any previous research. For IRT-grm, we expected a slight bias for skewed items loading on a normal LV (Expectation 4e), so these findings are slightly better than expected.

FA-lin underestimates the parameters ($-5.6\%$, $-12.1\%$, $-19.0\%$, and $-6.2\%$ RB, for normal, skewed, mixed-skewed, and bimodal items, respectively), which is in accordance with Expectation 2b for the skewed items. Thus, for FA-lin the loading estimator of a bimodal item loading on a normal LV is only slightly more biased than the loading estimator of a normally distributed item, whereas the loading estimator of skewed items loading on a normal LV is severely biased, and more so when both directions of skewness are present in a scale. It should be noted that the normal item presented here is taken from a scale with normal items only; the loading estimators

*Table 6.2.* MANOVA results: $\eta_p^2$ per effect for RB of $\lambda$ parameters and of corresponding standard errors. $N = 60000$.

| Effect | Levels | Parameters $\eta_p^2$ | Standard errors $\eta_p^2$ |
|---|---|---|---|
| Model (m) | 3 | 0.851 | 0.135 |
| LV distribution (lv) | 2 | 0.239 | |
| Scale shape (ss) | 5 | 0.501 | 0.085 |
| Sample size (n) | 2 | | |
| m × lv | 6 | 0.029 | 0.059 |
| m × ss | 15 | 0.444 | 0.101 |
| lv × ss | 10 | 0.376 | 0.089 |
| m × lv × ss | 30 | 0.122 | 0.074 |
| Item group (ig) | 6 | 0.483 | 0.224 |
| m × ig | 18 | 0.273 | 0.162 |
| lv × ig | 12 | 0.310 | 0.171 |
| ss × ig | 30 | 0.445 | 0.180 |
| m × lv × ig | 36 | 0.058 | 0.080 |
| m × ss × ig | 90 | 0.258 | 0.172 |
| lv × ss × ig | 60 | 0.392 | 0.193 |
| m × lv × ss × ig | 180 | 0.140 | 0.159 |

*Note.* Listed effects are statistically significant at $\alpha = 0.01$ and are sized $\eta_p^2 > 0.02$.

of normal items from the other scale types (not plotted here, but available from Tables E.1 to E.60 in Appendix E.1) are equally biased, with one exception: in the scale including normal items and items of one type of skewness, the loading estimator of the normal items are more severely biased (−7.7% RB).

Second, turning to the skew-normal LV results, we see that, of the three models, IRT-grm is least affected by the nonnormal LV distribution, with an unacceptable bias of loading estimators only of the skewed item loading on a LV of opposite skewness (−5.8% and −6.5% RB for the skewed and mixed-skewed items, respectively). These results support Expectations 4g and 4h.

For FA-lin the left-skewed and ls-mix items loading on the right skew-normal LV are also the most problematic, resulting in spectacular RBs of −29.6% and −38.1%, respectively. The cause of this extreme bias could be sought in the fact that the thresholds of these items are nowhere near being evenly spaced, which is a prerequisite for acceptable FA-lin parameter estimation results. For the normal and the bimodal items, the results also worsen, compared to the normal LV condition (−8.3% and −8.1% RB for normal and bimodal items, respectively), which is in line with Expectation 2d. The estimation of the loading parameters of the right-skewed items, however, *improves* for the right skew-normal LV compared to the normal LV, resulting in an acceptable RB level of −2.0%. This is due to the evenly spaced thresholds of the right-skewed items loading on a right skew-normal LV, and consistent with Expectation 2c. However, we did not expect the estimator to perform even better than in the normal condition.

FA-poly results are worse for the skewed LV than for the normal LV for every item distribution. The loadings for normal and bimodal items with a skew-normal LV are underestimated slightly, but within the acceptable range, so for the normal items these results are better than we tentatively expected (Expectation 3e). Loading estimators are positively biased for the right-skewed items loading on a right skew-normal LV (7.3% RB) and negatively biased for the left-skewed items (−12.8% RB). These findings are in accordance with Expectations 3f and 3g, respectively.

The precision of the estimators is higher for normal than for skewed item distributions, regardless of the LV distribution, which corresponds to Expectation 3k, and pinpoints the source of that effect to the item rather than the LV distribution.

When comparing the results from the small sample size with those of the medium sample size, it is clear that the precision of the loading estimators increases considerably with increasing sample size. The bias, however, is quite constant, as can also be concluded from the lack of significance or importance of the effects of sample size in the MANOVA.

**Standard Errors** The results from the MANOVA on the standard error estimates are presented in the last column of Table 6.2. Many significant but small effects are found, the most important of which is, as with the parameter estimators, the interaction of LV distribution, scale shape, model, and item group ($\eta_p^2 = 0.159$). The results are illustrated in Figures 6.3 and 6.4 (note the difference in scale of the vertical

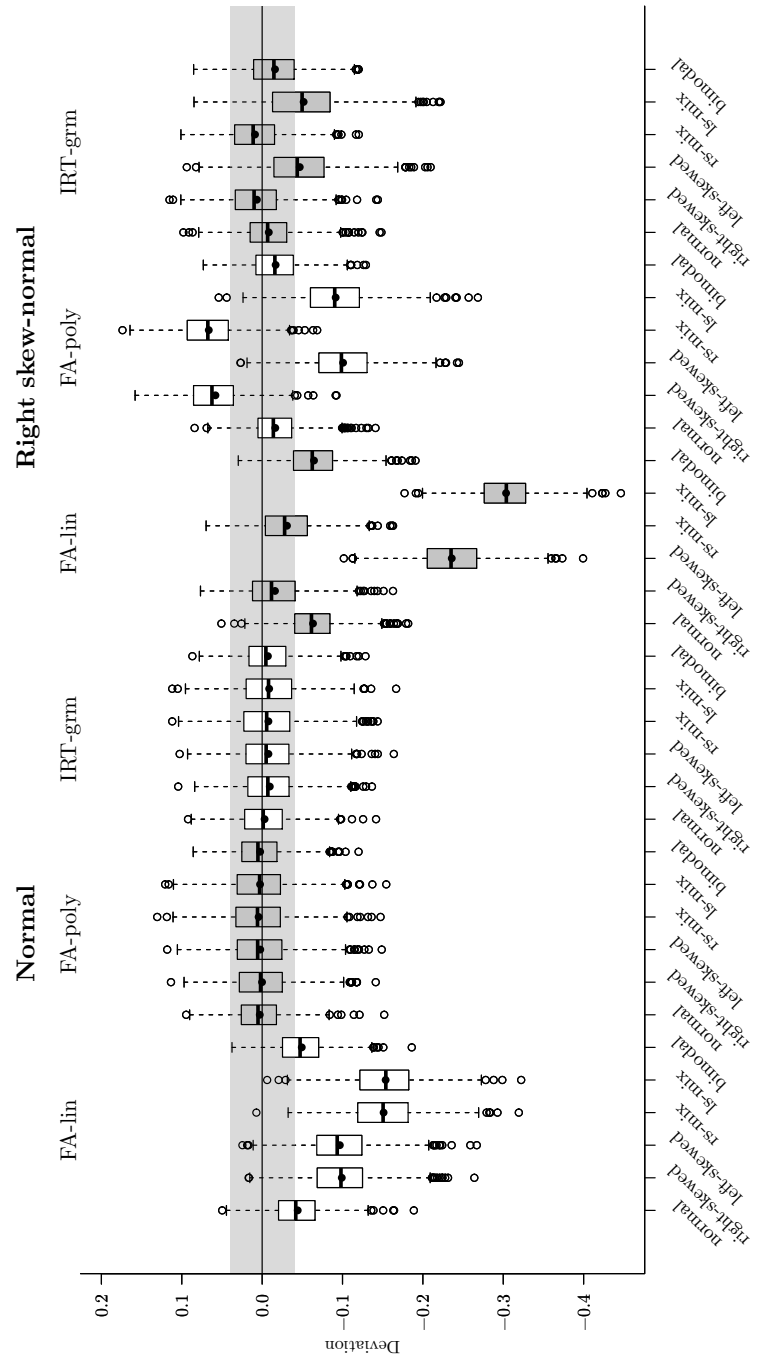



*Figure 6.1.* $\hat{\lambda}_i - \lambda_i$ for a normal and right skew-normal LV and various item types, as estimated by FA-lin, FA-poly, and IRT-grm. The grey area represents the margin of deviation around the true values considered acceptable. $n = 200$; $R = 1000$.

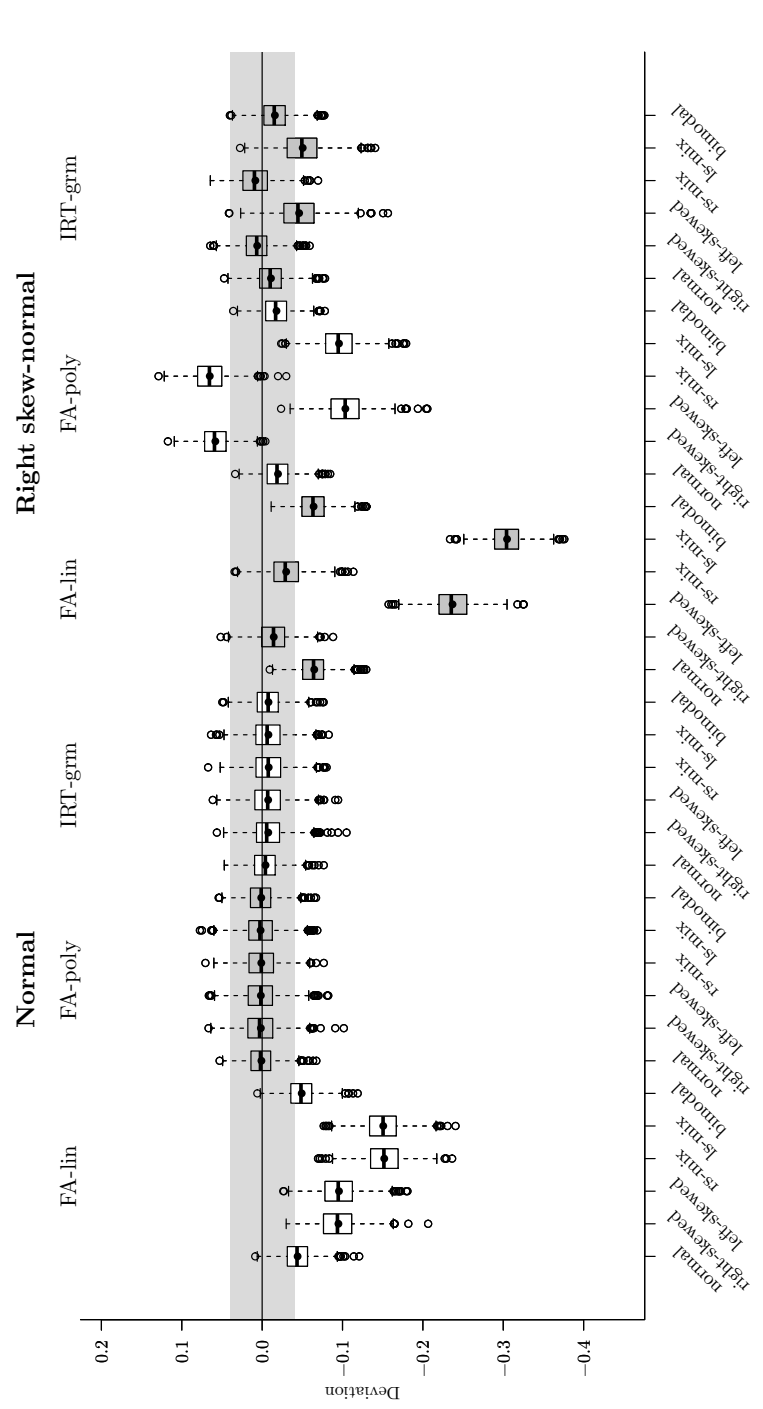



*Figure 6.2.* $\hat{\lambda}_i - \lambda_i$ for a normal and right skew-normal LV and various item types, as estimated by FA-lin, FA-poly, and IRT-grm. The grey area represents the margin of deviation around the true values considered acceptable. $n = 600$; $R = 1000$.

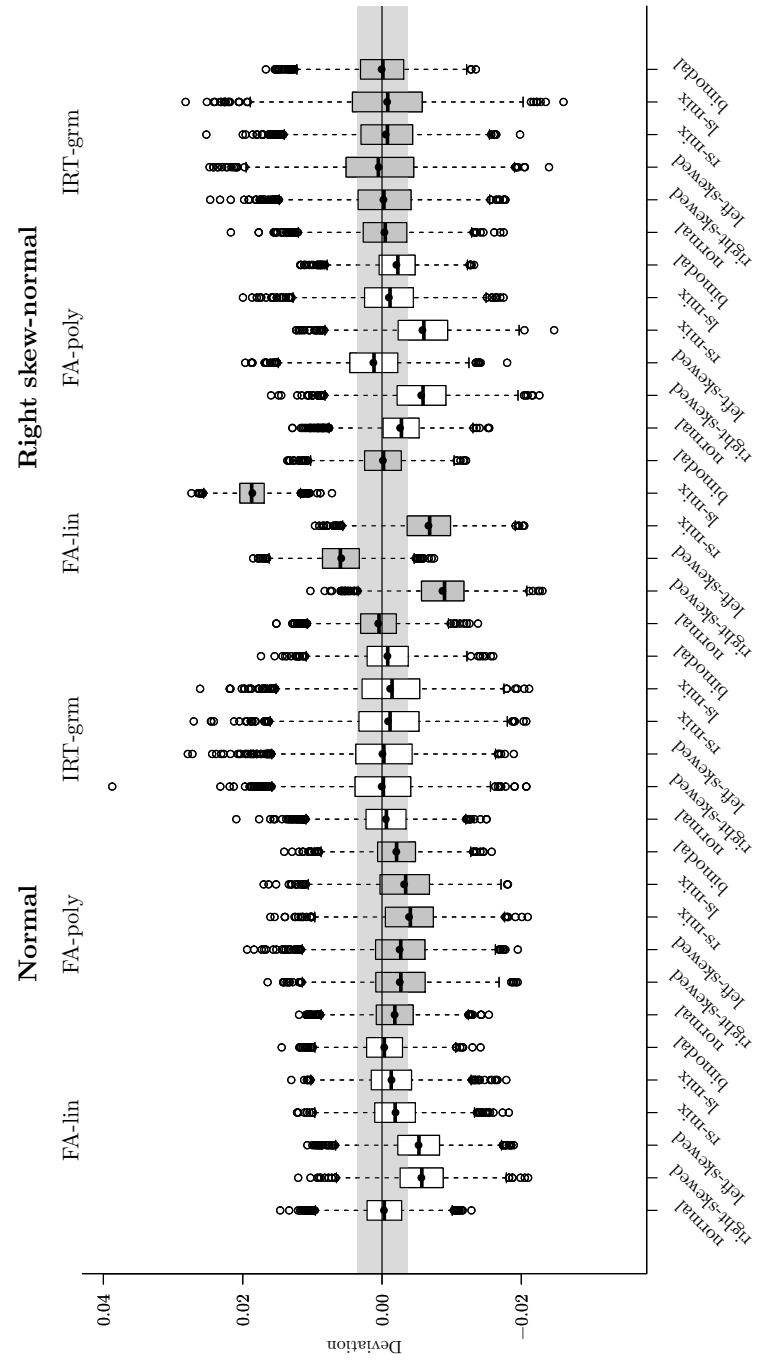



*Figure 6.3.* $\hat{se}(\hat{\lambda}_i) - sd(\hat{\lambda}_i)$ for a normal and right skew-normal LV and various item types, as estimated by FA-lin, FA-poly, and IRT-grm. The grey area represents an approximation to the margin of deviation around the true values considered acceptable. $n = 200$; $R = 4000$.

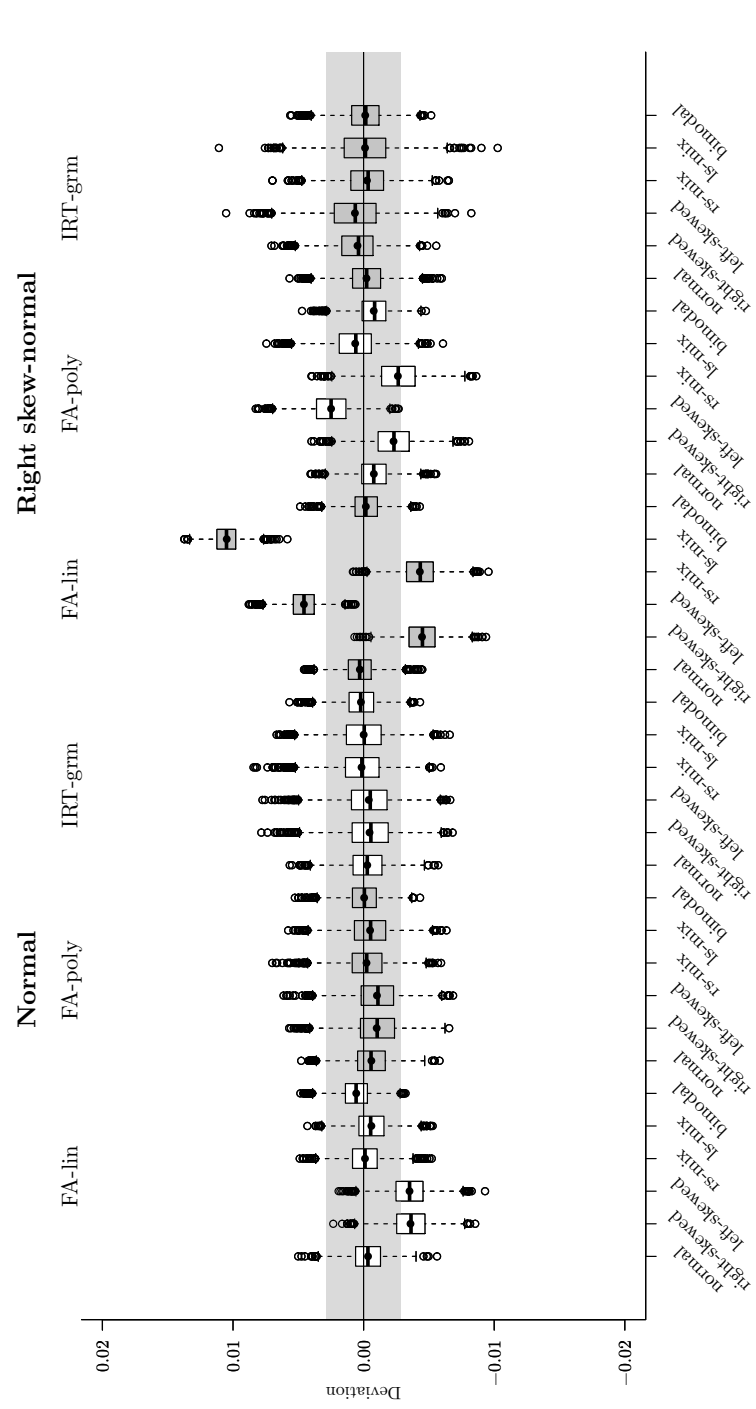



*Figure 6.4.* $\hat{se}(\hat{\lambda}_i) - sd(\hat{\lambda}_i)$ for a normal and right skew-normal LV and various item types, as estimated by FA-lin, FA-poly, and IRT-grm. The grey area represents an approximation to the margin of deviation around the true values considered acceptable. $n = 600$; $R = 4000$.

axis), showing the deviation of the standard error estimates from the parameter standard deviations for the small and medium sample size, respectively. Because the RB of standard error estimators for individual items of the same shape varied considerably, estimates are taken together for four items of equal shapes, resulting in 4000 replications per boxplot, rather than 1000.

We first turn to the normal LV conditions. IRT-grm results stand out by their lack of bias for any of the item types, which is in accordance with Expectation 4m. From the relatively large spread of the estimates we infer that the parameter estimators are relatively unprecise for the skewed items, compared to the other estimation models. This difference in precision is not expressed in a relatively high root mean squared error (RMSE) for IRT-grm compared to the other models (see, e.g., Tables E.19 to E.21), because for FA-lin and FA-poly the standard error estimators are more biased, and the RMSE is affected by both the bias and the precision of an estimator.

FA-lin standard error estimators are biased only for the right-skewed and left-skewed items (−12.9% RB), which partly supports our Expectation 2e. The results for the normal, bimodal, rs-mix, and ls-mix items are all well within our 10%-criterion of acceptable bias.

FA-poly standard errors of loading parameters are underestimated for each item distribution (mostly between −5% and −7% RB), but all within the 10%-criterion, which is in accordance with Expectation 3m.

Turning to the skew-normal LV conditions, the high accuracy but relatively low precision of the IRT-grm standard error estimators again becomes apparent: The standard error estimators are unbiased for every item distribution, which supports Expectation 4p, but also exhibit the largest variance of all estimation models.

FA-lin standard error estimators are unbiased for normal and bimodal items loading on a skew-normal LV. For skewed items, however, they are biased considerably with −21.8%, 14.4%, −17.9%, and 48.4% RB for the right-skewed, left-skewed, rs-mix, and ls-mix items, respectively, which is in accordance with Expectation 2e.

Remarkably, the standard errors of loading parameters of left-skewed and ls-mix items are the least biased for FA-poly. Standard errors are biased substantially for right-skewed and rs-mix items (−13.7%, −13.3% RB), but also for normal items in scales that include right-skewed items (RB between −8.1 and −12.6%; not presented in the figures, but see Tables E.38, E.41, E.50 and E.53 in Appendix E.1). The increased bias of right-skewed items loading on a right skew-normal LV supports Expectation 3p.

Comparing the small and medium sample size, we observe the same pattern of results with regard to bias, with a decreased magnitude of bias for increasing sample size. The precision of the loading standard error estimators clearly improves with an increasing sample size.

### Thresholds

**Parameters** For threshold parameters we are interested in the plain rather than the relative bias, because we regard a deviation from the population value equally

important regardless of whether it is at the end or in the middle of the latent scale. We therefore applied a MANOVA to the PB-constituents of the threshold parameter estimators. For the standard error estimators we applied a MANOVA to the RB-constituents.

In Table 6.3 the effect sizes resulting from these MANOVAs are given for the explanatory variables that significantly affected the response variables at a significance level of $\alpha = 0.01$ when $\eta_p^2 > 0.02$. Since threshold parameters are not included in the FA-lin model, the analyses were applied to FA-poly and IRT-grm results, which is reflected by the two levels of the explanatory variable model.

With regard to the PB of parameter estimators, we observe many significant effects. Although the main effect of threshold type is the largest ($\eta_p^2 = 0.680$), the most important effect is, presumably, the interaction of model, LV distribution, and threshold type ($\eta_p^2 = 0.503$). This implies that the FA-poly parameter estimators are more biased for the skew-normal than the normal LV distribution and more so for the outer than for the inner thresholds, which can be observed in Figures 6.5 and 6.6.

From those figures we infer that both models perform rather well in case of a normal LV, regardless of the item distribution, which is in accordance with Expectations 3d and 4f.

FA-poly estimators are biased (up to $-0.18$ PB) for the right skew-normal LV, and more so for the outer than for the inner thresholds. For the normal items, the outer thresholds are underestimated, which supports Expectation 3h the inner thresholds are overestimated, but only slightly. For the right-skewed and rs-mix items, all thresholds are biased towards the middle, i.e., the first two thresholds are underestimated and the second two thresholds are overestimated, which partly corresponds to Expectation 3i, where we expected overestimation of threshold parameter estimators, based on studies involving dichotomous items. For the left-skewed and ls-mix items, an almost reverse pattern can be observed: the first threshold is overestimated, the second is unbiased, and the third and fourth are underestimated. For the first two thresholds, the bias is less severe than in case of the right-skewed items; for the second two thresholds the bias is more severe than in case of the right-skewed items. These results partly support Expectation 3j, where we expected a more severe positive bias for the left-skewed than for the right-skewed items.

IRT-grm estimators, on the other hand, remain quite accurate, and are only affected in case of left-skewed or ls-mix items loading on a right skew-normal LV (up to $-0.07$ PB), which is in accordance with Expectation 4k. For the normal and right-skewed items, the threshold estimators are more accurate than we expected (Expectations 4i and 4j).

It seems that the threshold values, as estimated by FA-poly, are rather unaffected by the underlying LV distribution, as is inferred from the almost identical estimates for both LV distributions, resulting in good results for the normal LV conditions, but in quite some bias in case of a skew-normal LV. Furthermore, the variance of the threshold estimates increases with increasing absolute population values.

*Table 6.3.* MANOVA results: $\eta_p^2$ per effect for PB of $\tau$ parameters and RB of corresponding standard errors. $N = 40000$.

| Effect | Levels | Parameters $\eta_p^2$ | Standard errors $\eta_p^2$ |
|---|---|---|---|
| Model (m) | 2 | 0.024 | |
| LV distribution (lv) | 2 | 0.042 | |
| Scale shape (ss) | 5 | | 0.029 |
| Sample size (n) | 2 | | 0.074 |
| m × lv | 4 | 0.023 | |
| lv × ss | 10 | | 0.024 |
| lv × n | 4 | | 0.025 |
| ss × n | 10 | | 0.036 |
| lv × ss × n | 20 | | 0.058 |
| Item group (ig) | 6 | 0.064 | 0.097 |
| m × ig | 12 | 0.036 | |
| lv × ig | 12 | 0.073 | 0.064 |
| ss × ig | 30 | | 0.049 |
| m × lv × ig | 24 | 0.027 | |
| lv × ss × ig | 60 | | 0.035 |
| lv × n × ig | 24 | | 0.028 |
| ss × n × ig | 60 | | 0.036 |
| lv × ss × n × ig | 120 | | 0.027 |
| Threshold type (t) | 2 | 0.680 | 0.275 |
| m × t | 4 | 0.504 | |
| lv × t | 4 | 0.678 | |
| ss × t | 10 | 0.148 | 0.044 |
| n × t | 4 | | 0.091 |
| m × lv × t | 8 | 0.503 | |
| m × ss × t | 20 | 0.146 | |
| lv × ss × t | 20 | 0.142 | |
| ss × n × t | 20 | | 0.058 |
| m × lv × ss × t | 40 | 0.144 | |
| lv × ss × n × t | 40 | | 0.028 |
| ig × t | 12 | 0.358 | 0.057 |
| m × ig × t | 24 | 0.155 | |
| lv × ig × t | 24 | 0.349 | 0.035 |
| ss × ig × t | 60 | 0.073 | 0.037 |
| n × ig × t | 24 | | 0.040 |
| m × lv × ig × t | 48 | 0.161 | |
| m × ss × ig × t | 120 | 0.037 | |
| lv × ss × ig × t | 120 | 0.075 | 0.031 |
| lv × n × ig × t | 48 | | 0.022 |
| ss × n × ig × t | 120 | | 0.029 |
| m × lv × ss × ig × t | 240 | 0.030 | |
| lv × ss × n × ig × t | 240 | | 0.054 |

*Note.* Listed effects are statistically significant at $\alpha = 0.01$ and are sized $\eta_p^2 > 0.02$.

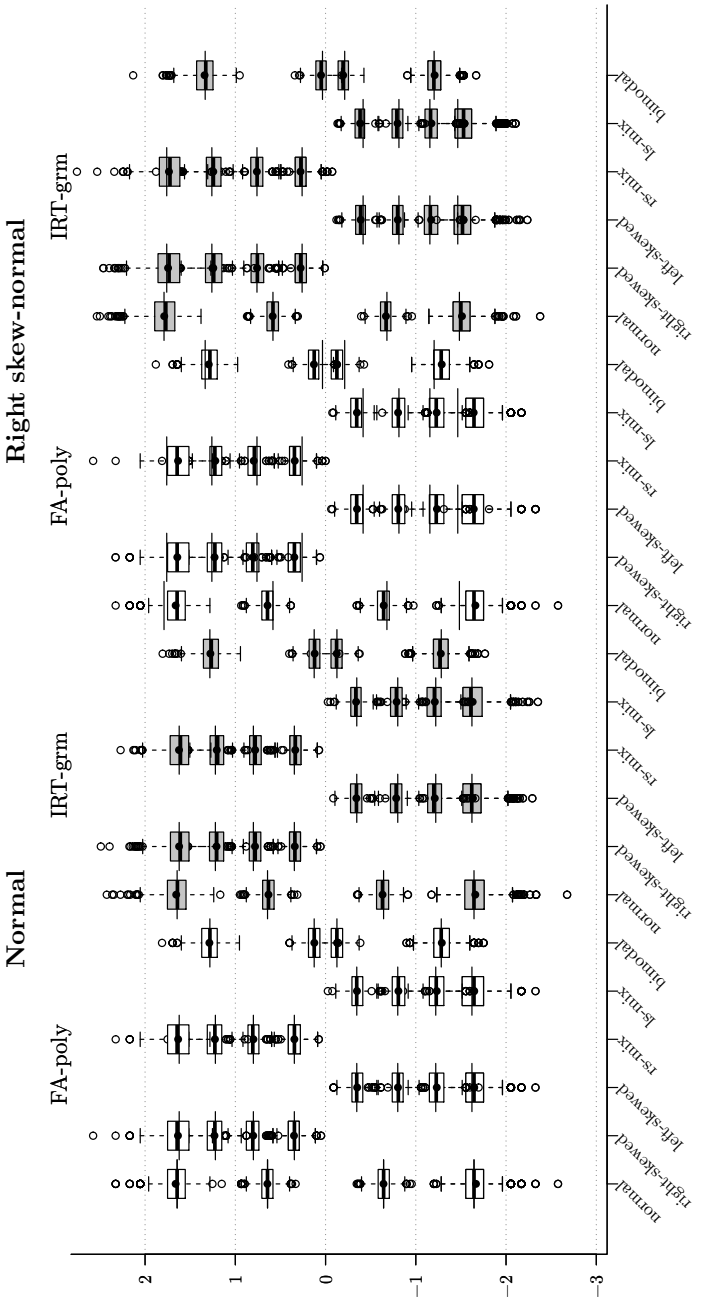



*Figure 6.5.* Parameter estimates $\hat{\tau}_{ic}$ for a normal and right skew-normal LV and various item types, as estimated by FA-poly and IRT-grm. The small horizontal lines crossing individual boxplots represent true $\tau_{ic}$ values. $n = 200$; $R = 1000$.

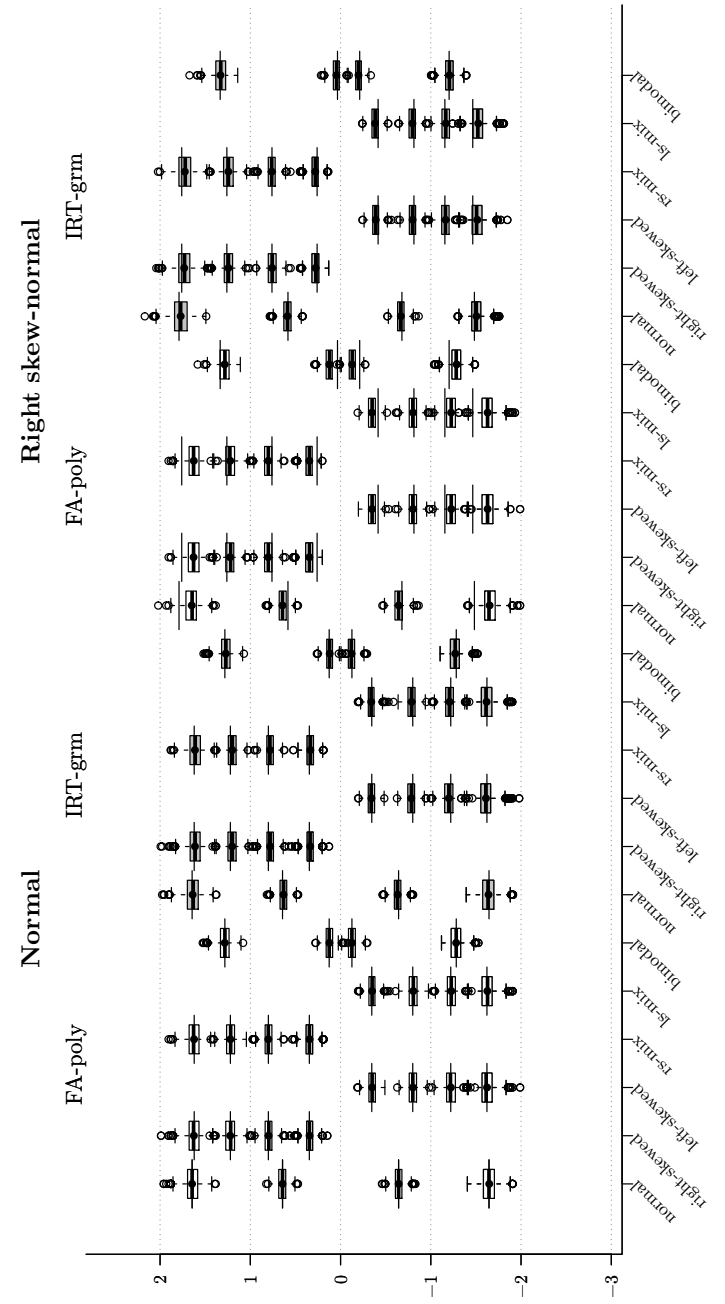



*Figure 6.6.* Parameter estimates $\hat{\tau}_{ic}$ for a normal and right skew-normal LV and various item types, as estimated by FA-poly and IRT-grm. The small horizontal lines crossing individual boxplots represent true $\tau_{ic}$ values. $n = 600$; $R = 1000$.

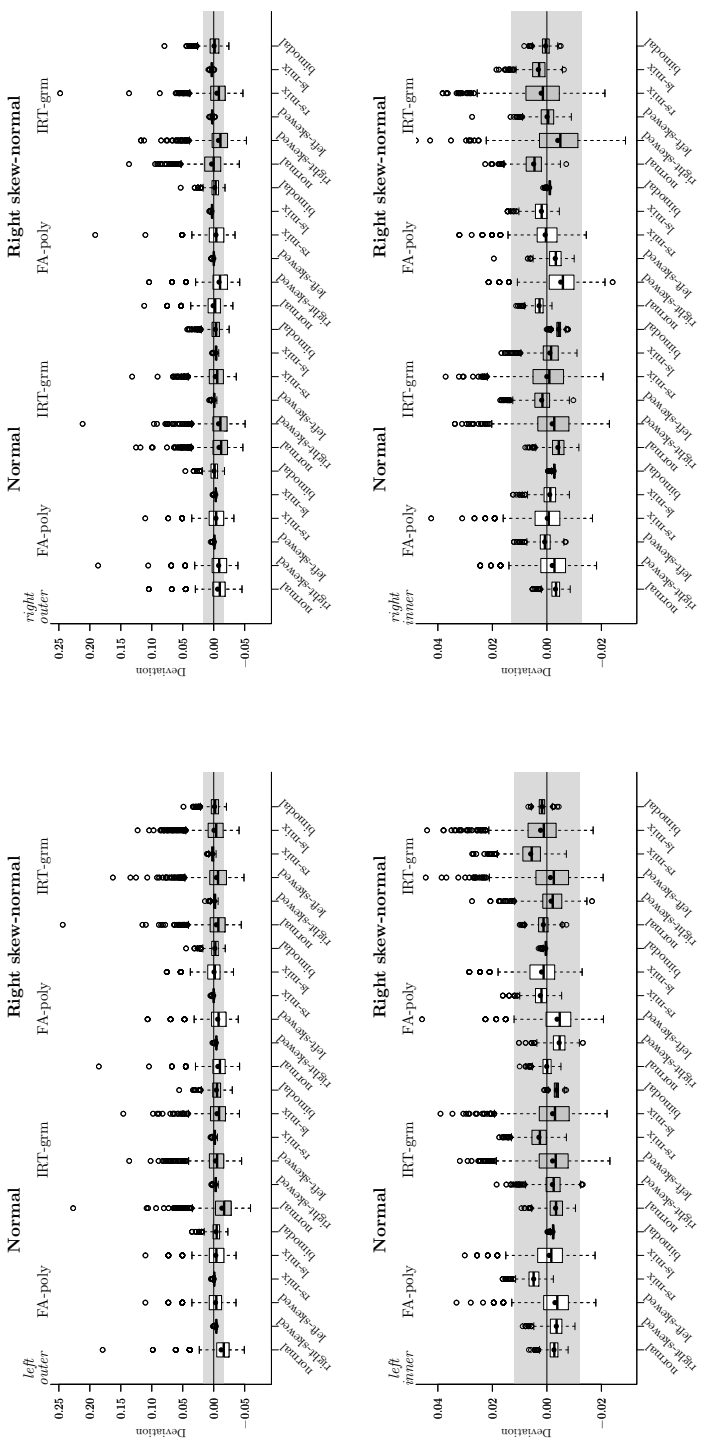



*Figure 6.7.* $\hat{se}(\hat{\tau}_{ic}) - sd(\hat{\tau}_{ic})$ for each of the four thresholds (see inset) for a normal and right skew-normal LV and various item types, as estimated by FA-poly and IRT-grm. The grey area represents an approximation to the margin of deviation considered acceptable. $n = 200$; $R = 1000$.

**Standard Errors** In the last column of Table 6.3 we find the MANOVA results for the RB of threshold standard errors, expressed as $\eta_p^2$. Many small significant effects are present, the largest of which is the main effect of threshold type ($\eta_p^2 = 0.275$), indicating better estimation of standard errors of the inner than of the outer threshold parameters.

Figure 6.7 presents the deviation of the threshold standard error estimates from the standard deviation of the parameter estimates for the small sample size. Notice the difference in scale between the upper and lower panels, representing the outer and inner thresholds, respectively. As with the figures of the loading standard errors, the grey area is an approximation to the deviation considered acceptable, based on the empirical standard deviation of the parameter estimates. Since the empirical standard deviations varied considerably between the item types, the maximum of the empirical standard deviations over item types was taken in each subfigure to approximate the margin of acceptable deviation.

For each condition, the distribution of standard error estimates is negatively skewed for both estimation models.

As the RB of threshold standard error estimators does not exceed our 10% criterion of acceptable deviation for any of the conditions for either estimation model, we conclude that both FA-poly and IRT-grm standard error estimation is good under conditions of nonnormal LV and/or item distributions, which is in accordance with Expectations 3o and 4o.

### Discrimination and Step-Difficulty Parameters

As was explained in Chapter 4, data were generated under the FA-poly model, although we could equivalently have used an IRT parameterization. To facilitate the comparison with IRT research, parameter and standard error estimation results of the IRT discrimination $\alpha$ and step-difficulty $\beta$ parameters are also presented, though more briefly than the FA-parameterized results. For the relations between the FA and IRT parameters we refer again to Section 4.1.2.

**Discrimination Parameters** In Table 6.4 results are presented of the MANOVA on discrimination parameters and their corresponding standard errors, as estimated by FA-poly and IRT-grm. Only two models are included in these analyses, as is apparent from the number of levels for the explanatory variable model. The results of the MANOVA applied to the discrimination parameter and standard error estimators resemble those of the loading parameter and standard error estimators (see Table 6.2), which is to be expected. The differences in results are mainly caused by the fact that for the loading parameters FA-lin was included in the analysis, whereas the comparison of discrimination parameter results only concerns FA-poly and IRT-grm. Furthermore, as discrimination parameters are on the logit scale, bias seems to be enlarged compared to the loading parameter bias.

The largest effect shown in Table 6.4 is the interaction between LV distribution, scale shape, and item group ($\eta_p^2 = 0.269$). Most other large effects are interaction

*Table 6.4.* MANOVA results: $\eta_p^2$ per effect for RB of $\alpha$ parameters and of corresponding standard errors. $N = 40000$.

| Effect | Levels | Parameters $\eta_p^2$ | Standard errors $\eta_p^2$ |
|---|---|---|---|
| Model (m) | 2 | 0.035 | 0.227 |
| LV distribution (lv) | 2 | 0.109 | |
| Scale shape (ss) | 5 | 0.264 | 0.103 |
| Sample size (n) | 2 | | 0.086 |
| m × ss | 10 | 0.123 | 0.094 |
| lv × ss | 10 | 0.254 | 0.105 |
| m × n | 4 | | 0.034 |
| m × lv × ss | 20 | 0.114 | 0.089 |
| Item group (ig) | 6 | 0.203 | |
| m × ig | 12 | 0.129 | |
| lv × ig | 12 | 0.217 | 0.024 |
| ss × ig | 30 | 0.267 | |
| m × lv × ig | 24 | 0.091 | |
| m × ss × ig | 60 | 0.098 | |
| lv × ss × ig | 60 | 0.269 | 0.027 |
| m × lv × ss × ig | 120 | 0.098 | |

*Note.* Listed effects are statistically significant at $\alpha = 0.01$ and are sized $\eta_p^2 > 0.02$.

or main effects involving these three variables. In addition, some smaller effects concerning model are found, the most important of which is the interaction of model, LV distribution, scale shape, and item group ($\eta_p^2 = 0.098$). From Figures 6.8 and 6.9 it can be observed that in case of a normal LV distribution, parameter estimation is good, regardless of the model or item distribution.

In case of the right skew-normal LV, the pattern of results is more diverse. For FA-poly all discrimination parameter estimators are severely biased for skewed items with 27.9%, −26.4%, 33.1%, and −24.6% RB for right-skewed, left-skewed, rs-mix, and ls-mix items, respectively, which supports Expectations 3f and 3g. For IRT-grm only the left-skewed and ls-mix items are biased considerably (−12.9% and −14.3% RB, respectively) supporting Expectation 4h. Right-skewed and rs-mix items are only slightly overestimated (3.4% and 4.3% RB, respectively), which is in accordance with Expectation 4g.

Discrimination parameters for normal and bimodal items are marginally biased for FA-poly (−5.3% and −5.0% RB, respectively) and IRT-grm (−2.4% and −4.5% RB, respectively).

The precision of the discrimination parameter estimators is rather poor for the small sample size, especially in case of right-skewed and rs-mix items loading on a right skew-normal LV for FA-poly, as is apparent from the long tails of the skewed distribution. For the medium sample size the precision of the estimators is improved considerably compared to the small sample size.

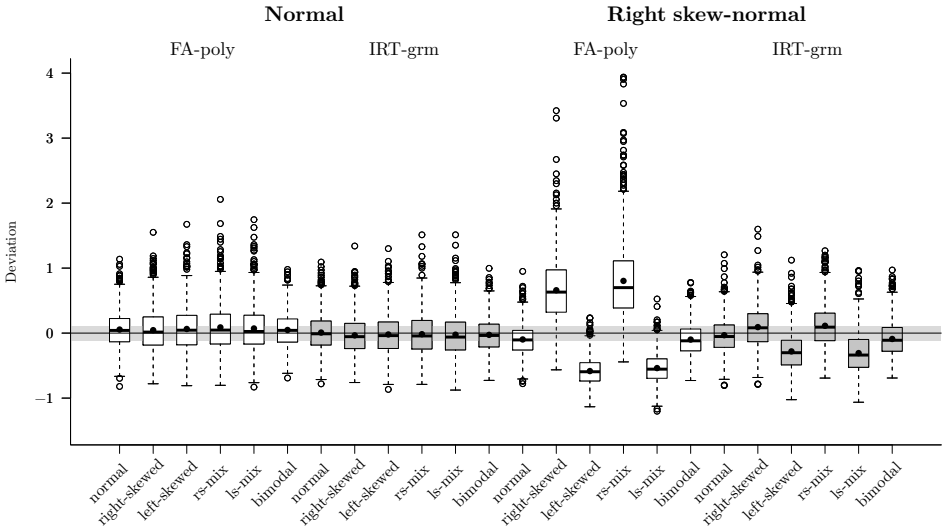



*Figure 6.8.* $\hat{\alpha}_i - \alpha_i$ for a normal and right skew-normal LV and various item types, as estimated by FA-poly and IRT-grm. The grey area represents the margin of deviation around the true values considered acceptable. $n = 200$; $R = 1000$.

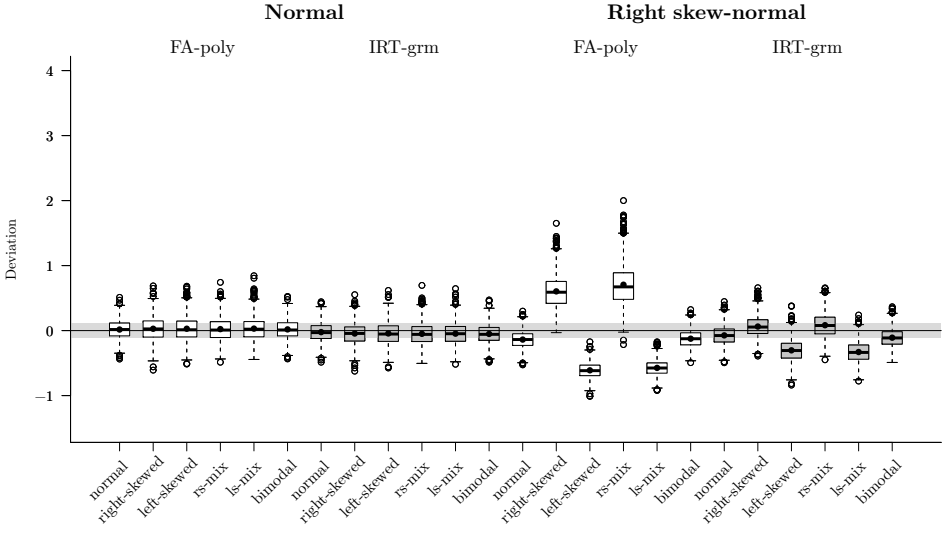



*Figure 6.9.* $\hat{\alpha}_i - \alpha_i$ for a normal and right skew-normal LV and various item types, as estimated by FA-poly and IRT-grm. The grey area represents the margin of deviation around the true values considered acceptable. $n = 600$; $R = 1000$.

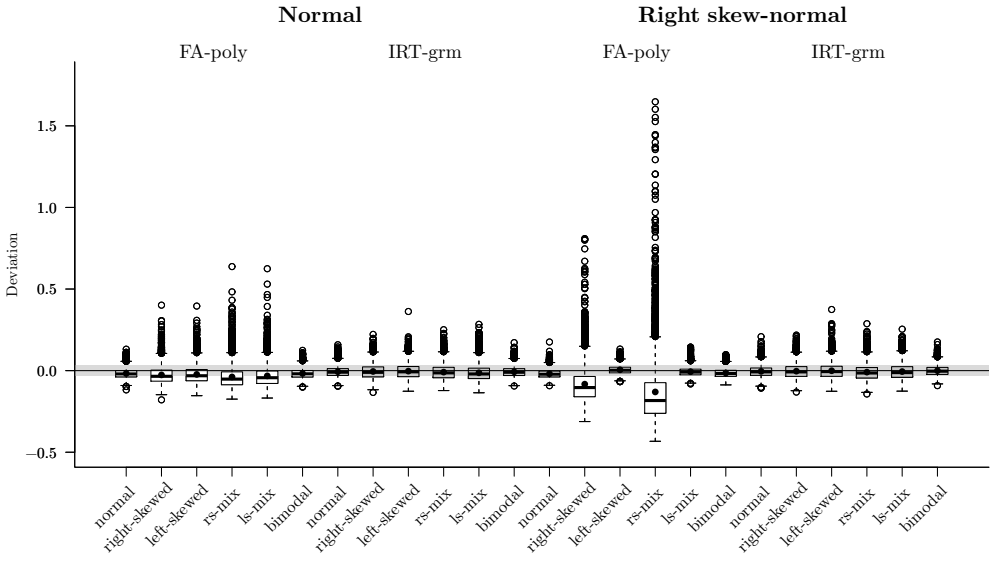



*Figure 6.10.* $\hat{se}(\hat{\alpha}_i) - sd(\hat{\alpha}_i)$ for a normal and right skew-normal LV and various item types, as estimated by FA-poly and IRT-grm. The grey area represents an approximation to the margin of deviation around the true values considered acceptable. $n = 200$; $R = 4000$.

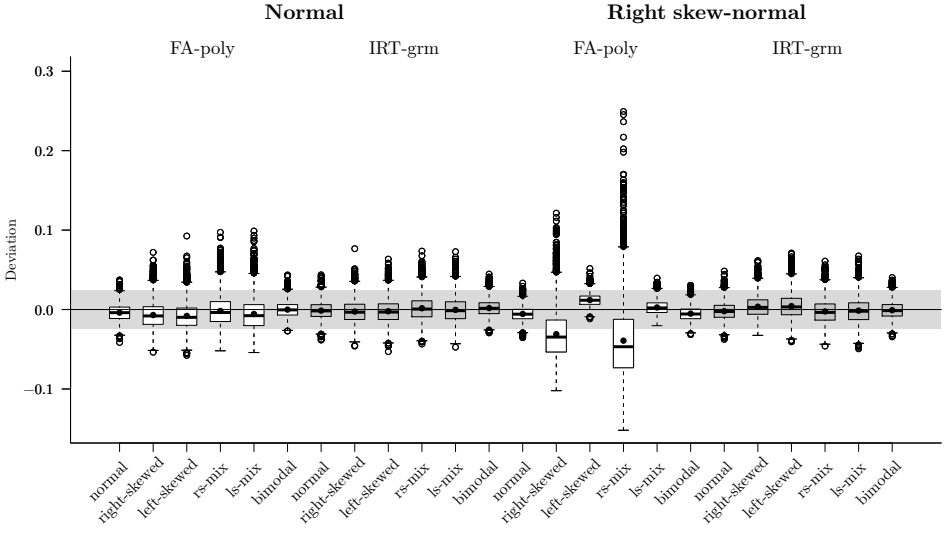



*Figure 6.11.* $\hat{se}(\hat{\alpha}_i) - sd(\hat{\alpha}_i)$ for a normal and right skew-normal LV and various item types, as estimated by FA-poly and IRT-grm. The grey area represents an approximation to the margin of deviation around the true values considered acceptable. $n = 600$; $R = 4000$.

**Discrimination Standard Errors** Discrimination standard error results are presented in the last column of Table 6.4. We observe a large main effect of model ($\eta_p^2 = 0.227$) and several medium and small effects, the most important of which is, arguably, the interaction of model, LV distribution, and scale shape ($\eta_p^2 = 0.089$).

In Figures 6.10 and 6.11 the deviation of the standard error estimates from the empirical standard deviation of the parameter estimates is depicted; note the difference in scales between the figures. For each item shape four items are taken together to accomodate for the considerable variation in RB of individual standard error estimators of equally shaped items, resulting in 4000 replications per boxplot rather than 1000.

It is clear that IRT-grm standard error estimators are, generally, unbiased, which supports Expectations 4m and 4p. FA-poly standard errors are moderately underestimated by about $-5$ to $-10\%$ for all item distributions in case of the normal LV, which is in accordance with Expectation 3m.

In case of the right skew-normal LV, FA-poly standard error estimators of right-skewed and rs-mix items stand out, as they are underestimated ($-15.1\%$ and $-16.4\%$ RB for right-skewed and rs-mix items, respectively), which is in accordance with Expectation 3p. Note that the distribution of estimates shows quite some variance and it is highly right-skewed. Standard error estimation bias decreases with increasing sample size, as does the variance of the distribution of estimates.

**Step-Difficulty Parameters** As the step-difficulty parameter estimation results were almost identical to those of the threshold parameters, they are not discussed separately. The interested reader is referred to Appendix E.2, Table E.85 and Figure E.1 for the MANOVA results and a corresponding graphical display, respectively.

### Coverage Rates

Because the coverage results very much resemble the parameter estimation results, they are discussed only briefly. Coverage rates of the 95%-confidence interval of loading and threshold parameter estimators for all parametric models are presented in Tables E.61 to E.64 in Appendix E.1.

One main result is that FA-lin coverage rates are unacceptable for every parameter in each cell.

FA-poly coverage rates are only affected by the LV distribution. When the LV distribution is normal, FA-poly coverage rates remain acceptable regardless of the item distribution. In the skew-normal LV conditions, FA-poly coverage rates are unacceptable for loading parameters of skewed items and for various threshold parameters in case of the small sample size. For the medium sample size, results are worse, with unacceptable coverage rates for almost every parameter estimator.

IRT-grm coverage rates are acceptable for all parameters in case of a normal LV distribution. When the LV distribution is skew-normal, coverage rates for most parameter estimators remain acceptable. In case of a small sample size, only loading parameter estimators of left-skewed items loading on the right skew-normal LV are

affected. For the medium sample size, the coverage rates of these parameters further deteriorate.

Coverage rates for (slightly) under- or overestimated parameters are worse for $n = 600$ than for $n = 200$, because we found parameter bias mostly to be stable over sample size, and standard errors to decrease with increasing sample size. This results in narrower confidence intervals, which, combined with even a small parameter bias, leads to worse coverage rates. This effect was also found in the normal data configurations discussed in Chapter 5.

### Summary of Parameter and Standard Error Results

**Parameters** Generally, loading parameter estimators are most biased for FA-lin, which is in accordance with Expectation 1b. The only unbiased FA-lin loading parameter is of a right-skewed item loading on a right skew-normal LV, which supports Expectation 2c. In accordance with our tentative Expectation 1c, FA-lin outperforms FA-poly in this condition. FA-poly and IRT-grm results are similar when the LV distribution is normal, with no loading parameter bias. In case of a skew-normal LV distribution, however, IRT-grm estimators outperform those of FA-poly, regardless of the item distribution, which supports Expectation 1d.

Similar findings apply to the threshold parameters, with unbiased estimators in case of the normal LV for both estimation models, and IRT-grm generally outperforming FA-poly in case of the skew-normal LV.

Discrimination parameter results much resemble those of the loading parameters. Effects are, however, somewhat enlarged as a result of the transformation of the loading paramter to the logit scale of the discrimination parameter. Step-difficulty parameter results are virtually equal to the threshold parameter, and are therefore not discussed separately.

All models perform notably well at estimating parameters of bimodal items. Results for bimodal items very much resemble those of normally distributed items. Perhaps distributional symmetry is the key factor of influence on parameter and standard error estimation.

**Standard errors** FA-lin loading standard errors are underestimated for right-skewed and left-skewed items loading on a normal LV, but only when no items of opposite skewness are present in the scale. In case of a skew-normal LV distribution, the loading standard error estimators of right-skewed and left-skewed items are severely negatively and positively biased, respectively, which supports Expectation 2e. As a result, significance testing of FA-lin parameter estimates will generally be too liberal in case of right-skewed items, and too conservative in case of left-skewed items.

FA-poly loading standard error estimators are unbiased in case of a normal LV, supporting Expectation 3m. When the LV distribution is right skew-normal, unacceptable negative bias is present for right-skewed items (in accordance with Expectation 3p) and for normal items in scales that include right-skewed items.

IRT-grm loading standard error estimators are unbiased in every condition (in accordance with Expectations 4m and 4p) but are notably the least efficient compared to the other estimation models.

Lack of efficiency is, in general, an issue for standard error estimators in case of the small sample size, impairing the reliability — and hence usefulness — of a standard error estimate.

Threshold standard error estimators do not deviate substantially from the empirical standard deviations for either FA-poly or IRT-grm in any of the conditions, and are thus considered robust, supporting Expectations 3o and 4o.

Discrimination standard error results are in accordance with the loading parameter results.

**Coverage** Coverage rates of the 95%-confidence intervals of parameter estimators are generally unacceptable for FA-lin. FA-poly coverage rates are acceptable when the LV is normal, but mostly unacceptable in case of the right skew-normal LV. IRT-grm coverage rates were the best, compared to FA-lin and FA-poly. Only in case of left-skewed items loading on the right skew-normal LV do we find unacceptable coverage rates for the loading parameter estimators.

**Summary** In conditions of a normal LV, both FA-poly and IRT-grm perform well at parameter and standard error estimation, regardless of the item distribution. When the LV is skew-normal, IRT-grm outperforms FA-poly, most significantly for right-skewed items and for normal items in scales including right-skewed items. FA-lin is outperformed by the other parametric models in all conditions, except in case of right-skewed items loading on a right skew-normal LV; FA-lin parameter estimators are unbiased then and more accurate than those of FA-poly.

### 6.2.3 Fit Indices

The effects of the explanatory variables on model fit are investigated by applying ANOVAs to the RMSEA based on the $\chi^2_{YB}$ and to the SRMR estimates, the results of which are presented in Table 6.5. The RMSEA results are provided in the second column. Although the main effect of model is the largest ($\eta^2 = 0.147$) for the RMSEA, the most important effect is the interaction between model and scale shape ($\eta^2 = 0.120$), indicating that for FA-lin the RMSEA is larger for scales that include skewed items, regardless of the LV distribution or the sample size.

In the last column the effect sizes resulting from the ANOVA on the SRMR estimates are listed. We observe two large main effects of scale shape ($\eta^2 = 0.216$) and of model ($\eta^2 = 0.211$). However, the most important effect is, presumably, the interaction of model, LV distribution, and scale shape ($\eta^2 = 0.063$), indicating the enlarged SRMR estimates for IRT-grm in case of the right skew-normal LV and skewed item distributions.

*Table 6.5.* ANOVA results: $\eta^2$ per effect for RMSEA and SRMR fit statistics. $N = 60000$.

| Effect | Levels | $\eta^2$ for RMSEA | $\eta^2$ for SRMR |
|---|---|---|---|
| Model (m) | 3 | 0.147 | 0.211 |
| LV distribution (lv) | 2 | | 0.095 |
| Scale shape (ss) | 5 | 0.055 | 0.216 |
| Sample size (n) | 2 | 0.034 | 0.065 |
| m × lv | 6 | | 0.145 |
| m × ss | 15 | 0.120 | 0.089 |
| lv × ss | 10 | | 0.040 |
| m × lv × ss | 30 | | 0.063 |

*Note.* Listed effects are statistically significant at $\alpha = 0.01$ and are sized $\eta^2 > 0.01$.

These effects can be inferred from Tables 6.6 and 6.7, providing the $\chi^2_{YB}$, RMSEA, and SRMR fit statistics averaged over replications for the small and medium sample size conditions, respectively, and for each estimation model. The degrees of freedom of the $\chi^2_{YB}$ distribution, which equals the expected value of the $\chi^2_{YB}$, are given in the table header for each model.

FA-lin fit statistics behave best in case of normal and bimodal items, which supports Expectation 2f for the normal items. Compared to the other models, FA-lin average $\chi^2_{YB}$ values show the largest deviation from the expected values. The deviation is largest for scales that contain skewed items, and increases with the number of skewed items included in the scale, which is in accordance with Expectation 2h, based on other $\chi^2$ statistics. The $\chi^2_{YB}$ had not been investigated in conditions of skewed items or a skewed LV before, but apparently it behaves in a similar way to other $\chi^2$ statistics then.

FA-poly and IRT-grm $\chi^2_{YB}$ values are very close to their expected value, supporting Expectations 3r and 4r. Consequently, the robustness of the $\chi^2_{YB}$ statistic to deviations from normality we tentatively expected for FA-lin and IRT-grm (Expectation 1g) is not found for FA-lin but is supported for IRT-grm. In addition, when applied to FA-poly, the $\chi^2_{YB}$ also turns out to be quite useful in conditions of nonnormality.

The SRMR results indicate a good fit in all cases for FA-lin and FA-poly, which conforms to Expectations 2j and 3s. Only for IRT-grm does the average SRMR reach or exceed the criterion of 0.08 in conditions where left-skewed items loading on a right skew-normal LV are included in the scale. It is remarkable that SRMR estimates for FA-lin and FA-poly are indicative of a closer model-data fit than IRT-grm SRMR estimates, even though IRT-grm seems to be superior over the other two models in terms of parameter estimators (see Section 6.2.2). Apparently, better parameter estimates do not necessarily lead to a model-implied covariance matrix that is more similar to the sample covariance matrix. An explanation for this finding might be that in FA-lin and FA-poly model estimation only univariate and bivariate information

*Table 6.6.* Fit statistics as estimated by the parametric models, averaged over replications. $n = 200$; $R = 1000$.

| Cell | Fit Statistic | FA-lin $df = 54$ | FA-poly $df = 18$ | IRT-grm $df = 18$ |
|---|---|---|---|---|
| nNS2 | $\chi^2_{YB}$ | 57.229 | 18.469 | 18.576 |
| | RMSEA | 0.016 | 0.017 | 0.018 |
| | SRMR | 0.026 | 0.026 | 0.032 |
| rnNS2 | $\chi^2_{YB}$ | 71.846 | 18.794 | 18.837 |
| | RMSEA | 0.039 | 0.018 | 0.018 |
| | SRMR | 0.039 | 0.031 | 0.041 |
| lnNS2 | $\chi^2_{YB}$ | 71.849 | 18.532 | 18.556 |
| | RMSEA | 0.039 | 0.017 | 0.017 |
| | SRMR | 0.039 | 0.031 | 0.041 |
| lrnNS2 | $\chi^2_{YB}$ | 82.429 | 18.874 | 18.908 |
| | RMSEA | 0.051 | 0.018 | 0.018 |
| | SRMR | 0.064 | 0.030 | 0.039 |
| bnNS2 | $\chi^2_{YB}$ | 59.088 | 18.843 | 18.995 |
| | RMSEA | 0.019 | 0.018 | 0.019 |
| | SRMR | 0.026 | 0.027 | 0.033 |
| nRS2 | $\chi^2_{YB}$ | 58.158 | 19.079 | 18.993 |
| | RMSEA | 0.018 | 0.020 | 0.019 |
| | SRMR | 0.028 | 0.029 | 0.044 |
| rnRS2 | $\chi^2_{YB}$ | 66.903 | 18.746 | 19.033 |
| | RMSEA | 0.032 | 0.018 | 0.019 |
| | SRMR | 0.031 | 0.035 | 0.077 |
| lnRS2 | $\chi^2_{YB}$ | 67.913 | 18.603 | 18.634 |
| | RMSEA | 0.033 | 0.017 | 0.017 |
| | SRMR | 0.044 | 0.044 | 0.097 |
| lrnRS2 | $\chi^2_{YB}$ | 78.616 | 18.577 | 18.637 |
| | RMSEA | 0.047 | 0.017 | 0.018 |
| | SRMR | 0.053 | 0.040 | 0.079 |
| bnRS2 | $\chi^2_{YB}$ | 58.599 | 18.645 | 18.671 |
| | RMSEA | 0.019 | 0.017 | 0.017 |
| | SRMR | 0.028 | 0.029 | 0.038 |

*Table 6.7.* Fit statistics as estimated by the parametric models, averaged over replications. $n = 600$; $R = 1000$.

| Cell | Fit Statistic | FA-lin $df = 54$ | FA-poly $df = 18$ | IRT-grm $df = 18$ |
|---|---|---|---|---|
| nNS6 | $\chi^2_{YB}$ | 55.790 | 18.361 | 18.304 |
| | RMSEA | 0.008 | 0.010 | 0.010 |
| | SRMR | 0.015 | 0.015 | 0.023 |
| rnNS6 | $\chi^2_{YB}$ | 99.346 | 18.167 | 18.239 |
| | RMSEA | 0.037 | 0.010 | 0.010 |
| | SRMR | 0.030 | 0.018 | 0.028 |
| lnNS6 | $\chi^2_{YB}$ | 98.703 | 18.467 | 18.449 |
| | RMSEA | 0.037 | 0.010 | 0.010 |
| | SRMR | 0.030 | 0.018 | 0.028 |
| lrnNS6 | $\chi^2_{YB}$ | 142.638 | 17.962 | 18.115 |
| | RMSEA | 0.052 | 0.009 | 0.009 |
| | SRMR | 0.060 | 0.017 | 0.027 |
| bnNS6 | $\chi^2_{YB}$ | 57.644 | 18.489 | 18.457 |
| | RMSEA | 0.010 | 0.010 | 0.010 |
| | SRMR | 0.015 | 0.015 | 0.023 |
| nRS6 | $\chi^2_{YB}$ | 56.001 | 17.979 | 18.079 |
| | RMSEA | 0.008 | 0.009 | 0.009 |
| | SRMR | 0.016 | 0.016 | 0.037 |
| rnRS6 | $\chi^2_{YB}$ | 82.943 | 17.821 | 18.121 |
| | RMSEA | 0.029 | 0.009 | 0.009 |
| | SRMR | 0.021 | 0.026 | 0.074 |
| lnRS6 | $\chi^2_{YB}$ | 84.841 | 18.341 | 18.424 |
| | RMSEA | 0.030 | 0.010 | 0.010 |
| | SRMR | 0.032 | 0.031 | 0.092 |
| lrnRS6 | $\chi^2_{YB}$ | 126.862 | 17.966 | 17.978 |
| | RMSEA | 0.047 | 0.009 | 0.009 |
| | SRMR | 0.047 | 0.029 | 0.075 |
| bnRS6 | $\chi^2_{YB}$ | 56.880 | 18.416 | 18.449 |
| | RMSEA | 0.009 | 0.010 | 0.010 |
| | SRMR | 0.016 | 0.017 | 0.029 |

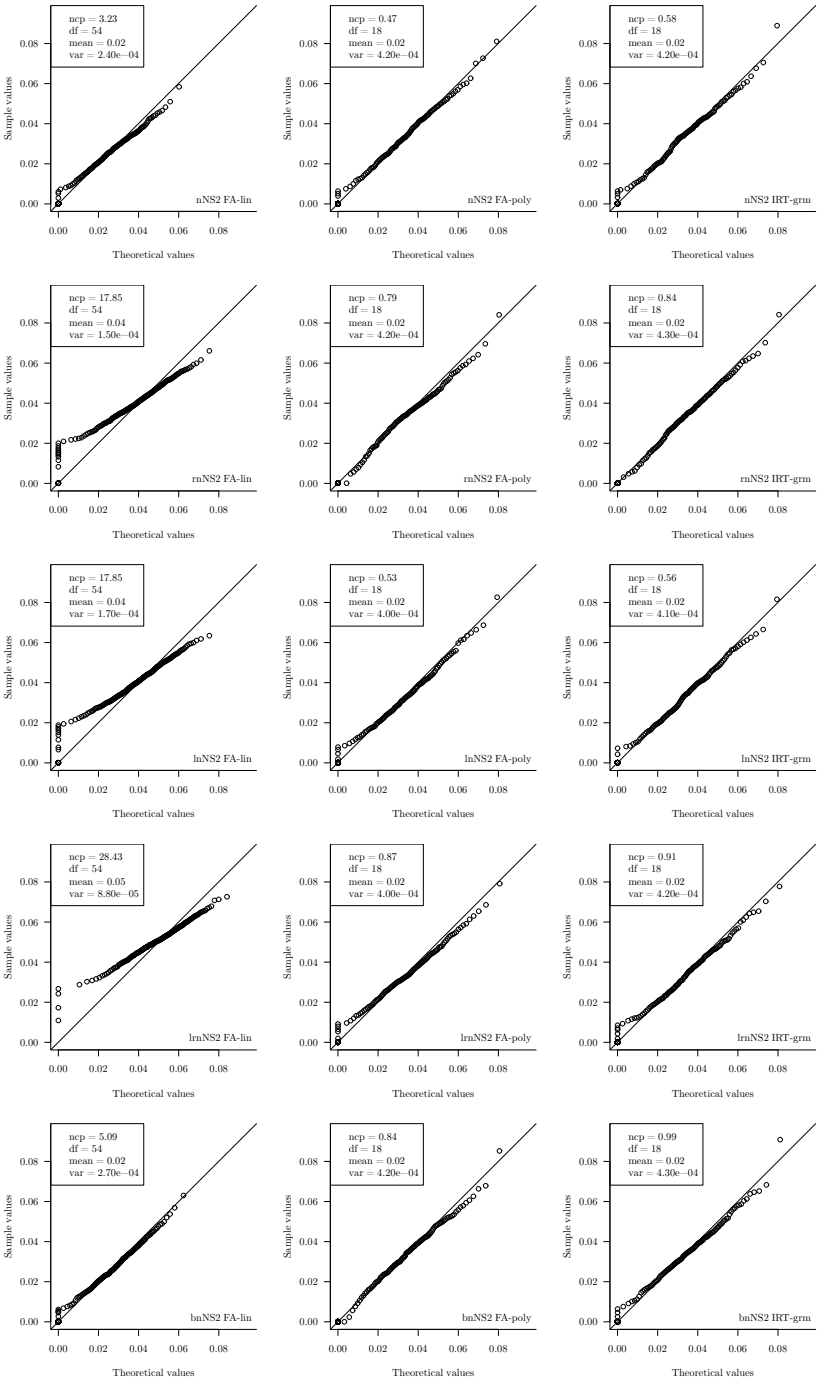



*Figure 6.12.* Q-Q plots for RMSEA fit statistic for Cells nNS2, rnNS2, lnNS2, lrnNS2, and bnNS2 and each model. LV distribution is normal. $n = 200$; $R = 1000$. The diagonal line depicts a perfect association between the empirical and theoretical distribution, the latter being a noncentral $\chi^2$ distribution using the mean empirical noncentrality parameter (NCP) over $R$ replications.

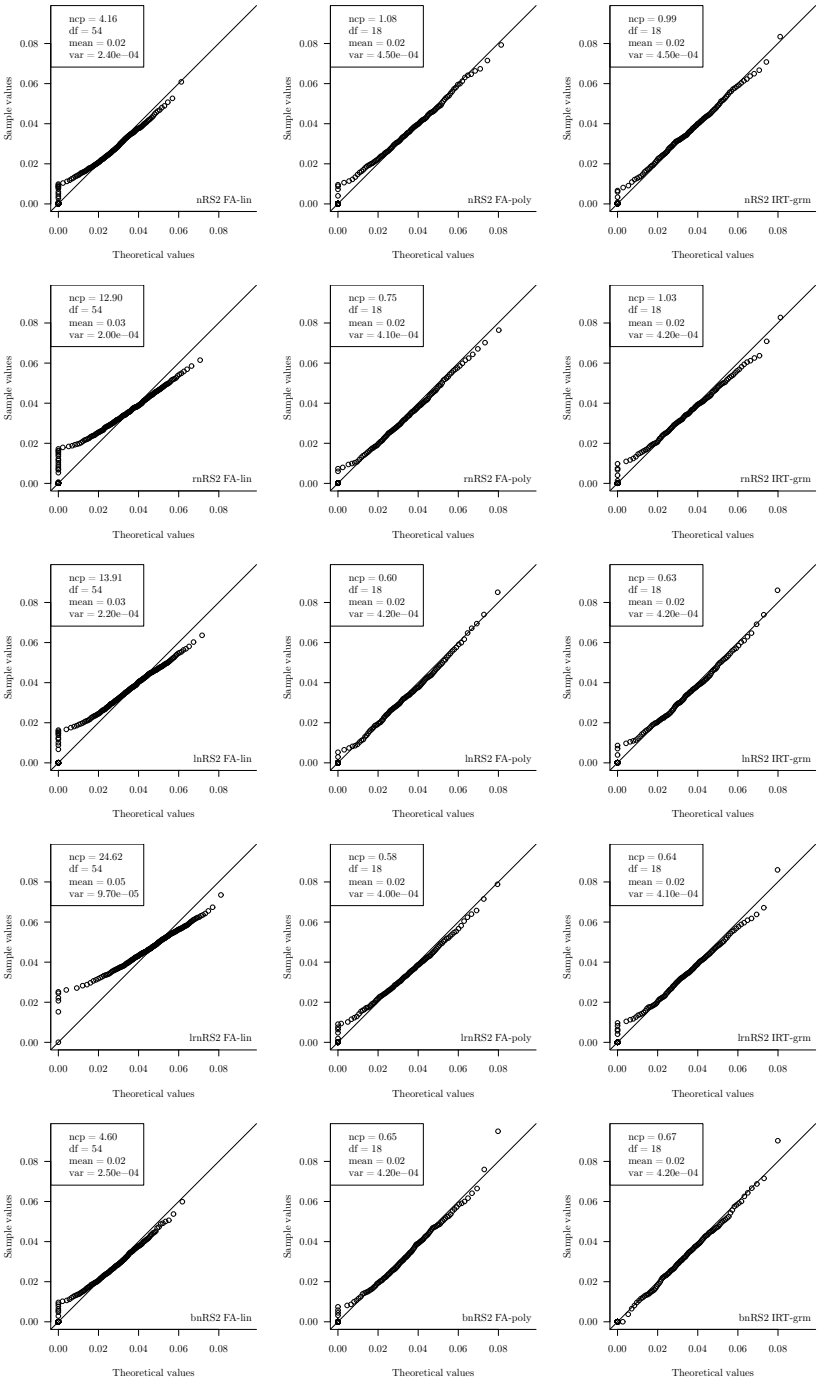



*Figure 6.13.* Q-Q plots for RMSEA fit statistic for Cells nRS2, rnRS2, lnRS2, lrnRS2, and bnRS2 and each model. LV distribution is right skew-normal. $n = 200$; $R = 1000$. The diagonal line depicts a perfect association between the empirical and theoretical distribution, the latter being a noncentral $\chi^2$ distribution using the mean empirical noncentrality parameter (NCP) over $R$ replications.

of the sample data is taken into account, whereas IRT-grm estimation makes use of the full response patterns. This difference between FA and IRT might explain why the bivariate sample information is more optimally fit by FA-lin and FA-poly than by IRT-grm. As the SRMR was not included in any previous Monte Carlo research including IRT-grm, these results are unprecedented.

The average RMSEA values never exceed the criterion of 0.06 for any of the models, including FA-lin. We are not only interested in average values, but in the entire distibution of the RMSEA statistic. Therefore, the distributions of the RMSEA are compared to their expected counterparts for the small sample size conditions involving the normal and skew-normal LV conditions in Figures 6.12 and 6.13, respectively. The Q-Q plots for the medium sample size conditions are given in Figures E.2 and E.3 in Appendix E.3.

We can observe the deviance of the FA-lin RMSEA values from the theoretical distribution when skewed items are involved. Clearly, the RMSEA as estimated by FA-lin is not suitable for such conditions. This result does not support our Expectation 2i, based on DiStefano (2002) who reported some robustness of FA-lin RMSEA against skewed items. However, the average RMSEA values do seem to support this hypothesis, and are indeed in line with DiStefano (2002) who did not provide any other information than the average RMSEA and 95%-confidence intervals for the population value. So our findings should be seen as an addition rather than a contradiction to that previous research.

RMSEA values as estimated by FA-poly and IRT-grm deviate only minimally from their theoretical counterparts, and do seem adequate for scales involving skewed item or LV distributions. This is in accordance with Expectations 3s and 4s. Our results indicate that the $\chi^2_{YB}$ statistic and the $\chi^2_{YB}$-based RMSEA are very useful for both FA-poly and IRT-grm modeling in normal as well as nonnormal distributional conditions.

### 6.2.4 Nonparametric IRT-mok

In this subsection the results of applying the nonparametric IRT-mok model are presented. We focus on the estimation of the scalability coefficient Loevinger's $H$ and corresponding standard errors, on item and scale level. The bias and other performance variables with regard to these parameters and standard errors can be found in Tables E.65 to E.84 of Appendix E.1.

Just as we did in the previous chapter, we first present the population $H$ values we derived from the loading and threshold configuration, which served as the basis of our data configurations.

The true $H$ values, serving as indicators of item and scale strength, are displayed in Table 6.8. Interestingly, items of heterogeneously shaped scales (consisting of items of various distributions) have larger $H$ values than items in homogeneous scales. These larger $H$ values are a result of the variation in item *means*, that occurs when items are variously shaped within a scale.

When comparing the true values in the normal LV conditions with those in the skew-normal LV conditions, all $H$ values are larger for the normal LV, except when the

*Table 6.8.* IRT-mok $H_i$ and $H_{scale}$ true values per cell of the design.

| | nNS2<br>nNS6 | rnNS2<br>rnNS6 | lnNS2<br>lnNS6 | lrnNS2<br>lrnNS6 | bnNS2<br>bnNS6 | nRS2<br>nRS6 | rnRS2<br>rnRS6 | lnRS2<br>lnRS6 | lrnRS2<br>lrnRS6 | bnRS2<br>bnRS6 |
|---|---|---|---|---|---|---|---|---|---|---|
| $H_1$ | n | n | n | n | n | n | n | n | n | n |
| | 0.571 | 0.604 | 0.604 | 0.615 | 0.633 | 0.542 | 0.630 | 0.517 | 0.584 | 0.603 |
| $H_2$ | n | n | n | n | n | n | n | n | n | n |
| | 0.571 | 0.604 | 0.604 | 0.615 | 0.633 | 0.542 | 0.630 | 0.517 | 0.584 | 0.603 |
| $H_3$ | n | n | n | n | n | n | n | n | n | n |
| | 0.571 | 0.604 | 0.604 | 0.615 | 0.633 | 0.542 | 0.630 | 0.517 | 0.584 | 0.603 |
| $H_4$ | n | n | n | n | n | n | n | n | n | n |
| | 0.571 | 0.604 | 0.604 | 0.615 | 0.633 | 0.542 | 0.630 | 0.517 | 0.584 | 0.603 |
| $H_5$ | n | n | n | l | n | n | n | n | l | n |
| | 0.571 | 0.604 | 0.604 | 0.635 | 0.633 | 0.542 | 0.630 | 0.517 | 0.512 | 0.603 |
| $H_6$ | n | n | n | l | n | n | n | n | l | n |
| | 0.571 | 0.604 | 0.604 | 0.635 | 0.633 | 0.542 | 0.630 | 0.517 | 0.512 | 0.603 |
| $H_7$ | n | r | l | l | b | n | r | l | l | b |
| | 0.571 | 0.572 | 0.572 | 0.635 | 0.614 | 0.542 | 0.660 | 0.417 | 0.512 | 0.588 |
| $H_8$ | n | r | l | l | b | n | r | l | l | b |
| | 0.571 | 0.572 | 0.572 | 0.635 | 0.614 | 0.542 | 0.660 | 0.417 | 0.512 | 0.588 |
| $H_9$ | n | r | l | r | b | n | r | l | r | b |
| | 0.571 | 0.572 | 0.572 | 0.635 | 0.614 | 0.542 | 0.660 | 0.417 | 0.699 | 0.588 |
| $H_{10}$ | n | r | l | r | b | n | r | l | r | b |
| | 0.571 | 0.572 | 0.572 | 0.635 | 0.614 | 0.542 | 0.660 | 0.417 | 0.699 | 0.588 |
| $H_{11}$ | n | r | l | r | b | n | r | l | r | b |
| | 0.571 | 0.572 | 0.572 | 0.635 | 0.614 | 0.542 | 0.660 | 0.417 | 0.699 | 0.588 |
| $H_{12}$ | n | r | l | r | b | n | r | l | r | b |
| | 0.571 | 0.572 | 0.572 | 0.635 | 0.614 | 0.542 | 0.660 | 0.417 | 0.699 | 0.588 |
| $H_{scale}$ | 0.571 | 0.586 | 0.586 | 0.629 | 0.623 | 0.542 | 0.647 | 0.460 | 0.599 | 0.595 |

*Note:* Item shape is indicated before each $H_i$ value, with n = normal, r = right-skewed, l = left-skewed, and b = bimodal.

*Table 6.9.* MANOVA results: $\eta_p^2$ per effect for RB of $H_i$ parameters and of corresponding standard errors. $N = 20000$.

| Effect | Levels | Parameters $\eta_p^2$ | Standard errors $\eta_p^2$ |
|---|---|---|---|
| LV distribution (lv) | 2 | | |
| Scale shape (ss) | 5 | 0.058 | 0.041 |
| Sample size (n) | 2 | | |
| ss × n | 10 | | 0.025 |
| lv × ss × n | 20 | | 0.041 |
| Item group (ig) | 6 | | 0.026 |
| lv × ig | 12 | | 0.021 |
| ss × n × ig | 60 | | 0.021 |
| lv × ss × n × ig | 120 | | 0.026 |

*Note.* Listed effects are statistically significant at $\alpha = 0.01$ and are sized $\eta_p^2 > 0.02$.

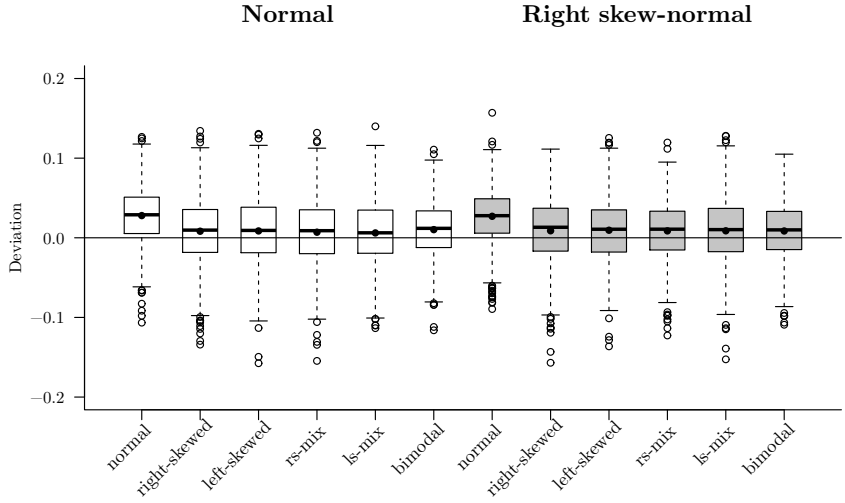



*Figure 6.14.* $\hat{H}_i - H_i$ for normal and right skew-normal LV conditions. $n = 200$; $R = 1000$.

item distribution is right-skewed, in which case the right skew-normal LV condition results in larger $H$ values.

**Parameters** In Table 6.9 the effect sizes resulting from the MANOVA applied to the RB-constituents of the $H_i$ parameter and standard error estimators are presented. Only scale shape substantially affects the bias of the parameter estimators ($\eta_p^2 = 0.058$). This is illustrated by Figures 6.14 and 6.15, showing the nonrelative deviation $\hat{H}_i - H_i$ for each item type in case of a small and medium sample size, respectively. Because the population $H$ values differ between the item types, it is not possible to mark the area of 5% deviation considered acceptable in the figures.

The largest, though still small, positive bias can be observed for the normal items (that belong to the all-normal-item scales). The distribution of parameter estimators is right-skewed and the precision is higher for the normal and the bimodal items as compared to the skewed items. Precision notably increases with increasing sample size.

For the $H_{scale}$ results, plotted in Figures 6.16 and 6.17 for the small and medium sample size, respectively, a similar pattern can be observed.

Parameter bias does not exceed our criterion set at 5% RB, and is thus considered acceptable in all conditions.

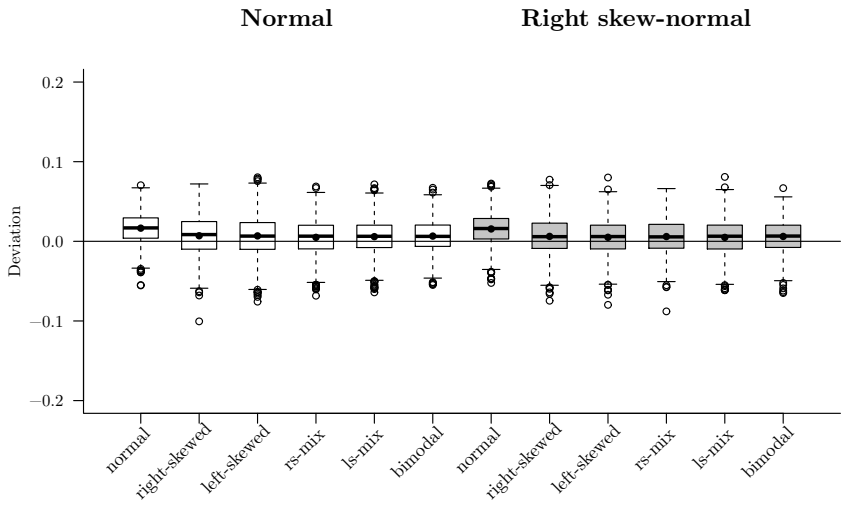



*Figure 6.15.* $\hat{H}_i - H_i$ for normal and right skew-normal LV conditions. $n = 600$; $R = 1000$.

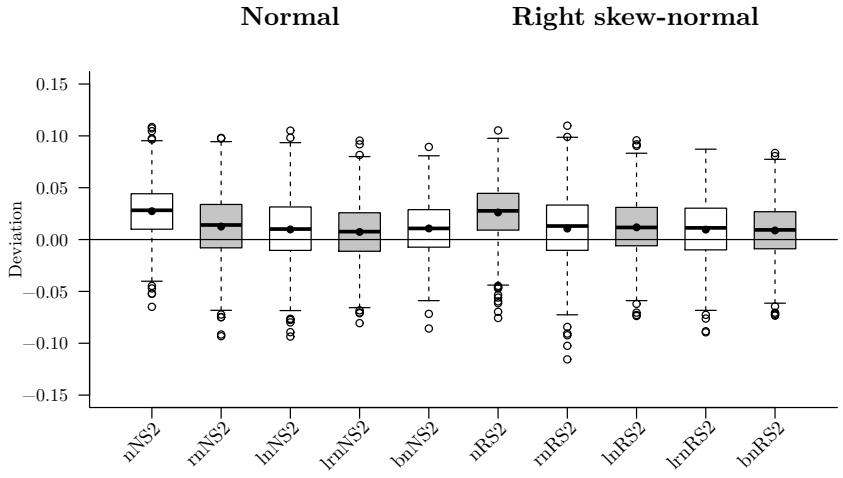



*Figure 6.16.* $\hat{H}_{scale} - H_{scale}$ for normal and right skew-normal LV conditions. $n = 200$; $R = 1000$.

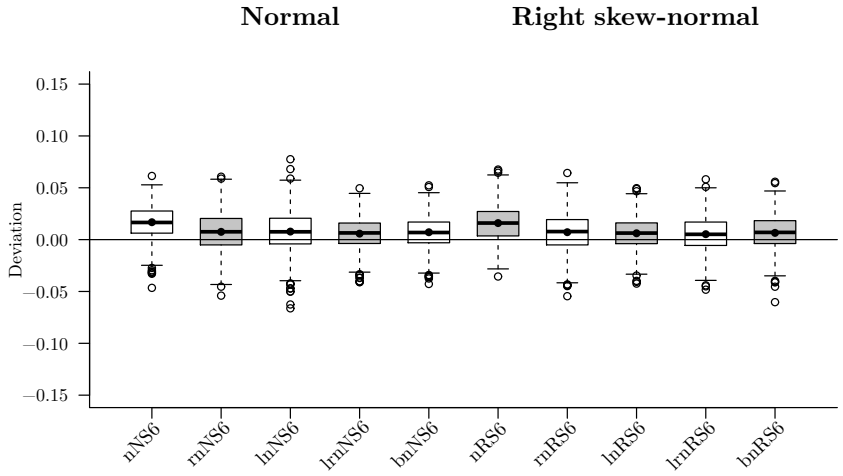



*Figure 6.17.* $\hat{H}_{scale} - H_{scale}$ for normal and right skew-normal LV conditions. $n = 600$; $R = 1000$.

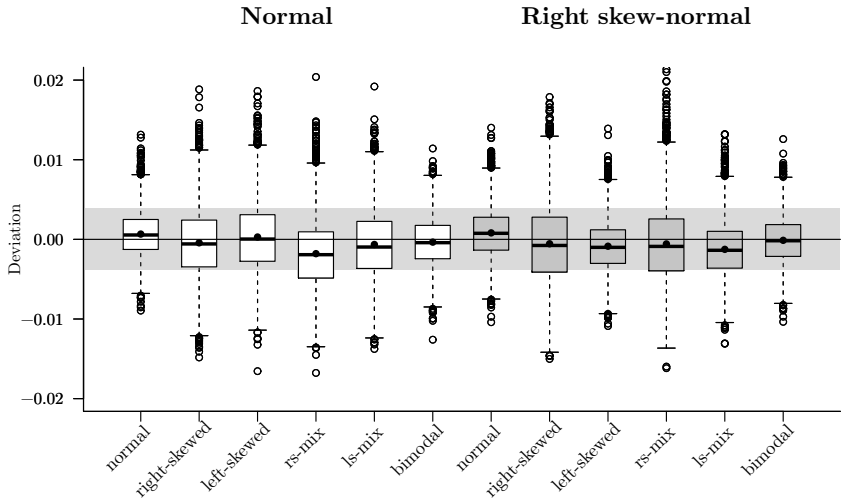



*Figure 6.18.* $\hat{se}(\hat{H}_i) - sd(\hat{H}_i)$ for normal and right skew-normal LV conditions. The grey area represents an approximation to the margin of deviation considered acceptable. $n = 200$; $R = 4000$.

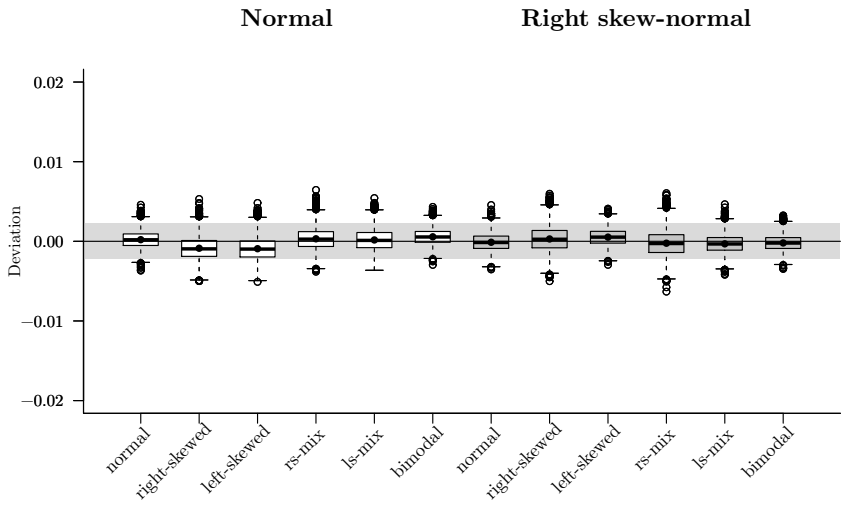



*Figure 6.19.* $\hat{se}(\hat{H}_i) - sd(\hat{H}_i)$ for normal and right skew-normal LV conditions. The grey area represents an approximation to the margin of deviation considered acceptable. $n = 600$; $R = 4000$.

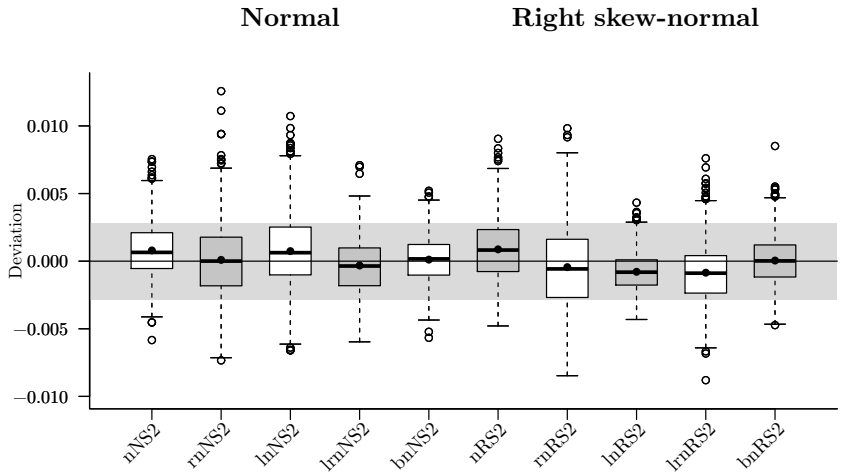



*Figure 6.20.* $\hat{se}(\hat{H}_{scale}) - sd(\hat{H}_{scale})$ for normal and right skew-normal LV conditions. The grey area represents an approximation to the margin of deviation considered acceptable. $n = 200$; $R = 1000$.

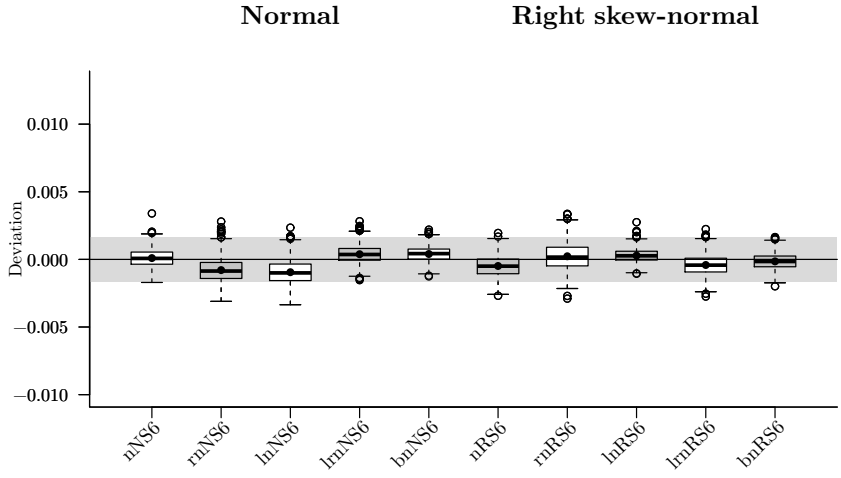



*Figure 6.21.* $\hat{se}(\hat{H}_{scale}) - sd(\hat{H}_{scale})$ for normal and right skew-normal LV conditions. The grey area represents an approximation to the margin of deviation considered acceptable. $n = 600$; $R = 1000$.

**Standard Errors and Coverage** The results of the MANOVA applied to the RB-constituents of the standard error estimators are reported in the last column of Table 6.9, where we see a number of significant but small effects, which do not appear to be substantially meaningful when regarding the graphical display of the standard error estimates in Figures 6.18 and 6.19 for the small and medium sample size, respectively.

Standard error estimates of $H_{scale}$ are depicted in Figures 6.20 and 6.21 for both respective sample sizes. Neither $H_i$ nor $H_{scale}$ standard error estimators are biased, as their RB is well within our acceptable range of 10% deviation.

In general, coverage rates of the $H_i$ and $H_{scale}$ parameters are a little low but acceptable at a rough average of 0.92 (see Appendix E.1.3). A notable exception are the conditions where the scale includes normal items only; coverage rates range between 0.80 and 0.87 then.

**Conclusion** The nonnormal LV and item distributions do not pose any problems for IRT-mok. On the contrary, scales with heterogeneously shaped item distributions lead to increased $H$ values that are estimated with less bias. Therefore, IRT-mok is very well suited for data from nonnormal LV and item distributions, which supports Expectations 1f and 5a. In addition, heterogeneous scales, with variously shaped items, result in higher Loevinger $H$ values compared to homogeneous scales.

### 6.2.5 Latent Variable Score Estimates

LV scores are estimated based on the estimated parameters, and thus differ between our scaling models. For IRT-mok LV scores are estimated simply by taking the unweighted sum scores over items in the scale.

#### Skewness of the Estimated LV distribution

As we are interested in how well the shape of the estimated LV distribution resembles the shape of the sampled LV distribution, we evaluate the deviation between the skewness of the estimated LV distribution and the skewness of the sampled true LV distribution. Results from the ANOVA applied to the deviation of skewness of LV scores are presented in the third column of Table 6.10.

Although the main effect of scale shape is the largest ($\eta^2 = 0.372$), the most interesting effects are, presumably, the interaction between model and scale shape ($\eta^2 = 0.320$) and the interaction between model, LV distribution, and scale shape ($\eta^2 = 0.015$), since the main factor scale shape is also involved in these interaction effects. These results are graphically displayed in Figures 6.22 and 6.23, where the deviation of the skewness of the estimated LV scores from the skewness of sampled true LV scores are provided for all four estimation models in each cell of the design.

First, turning to the results for the normal LV conditions, we find that the skewness of the distribution of LV scores as estimated by FA-poly and IRT-grm approximates the skewness of the sampled LV scores very well. For both models, there is a slight effect of the item distributions: When the scale consists of right-skewed and normal items, the LV distribution is estimated to be slightly right-skewed. For scales consisting of normal and left-skewed items, the distribution of the estimated LV scores is slightly left-skewed.

FA-lin and IRT-mok produce a different pattern of results. The distribution of estimated LV scores resembles the true distribution in case of scales consisting of

*Table 6.10.* ANOVA results: $\eta^2$ per effect for LV results. $N = 80000$.

| Effect | Levels | $\eta^2$ for Skewness deviation | $\eta^2$ for Kendall's $\tau_a$ |
|---|---|---|---|
| Model (m) | 4 | 0.014 | 0.095 |
| LV distribution (lv) | 2 | 0.195 | 0.262 |
| Scale shape (ss) | 5 | 0.372 | 0.200 |
| Sample size (n) | 2 | | |
| m × lv | 8 | 0.014 | |
| m × ss | 20 | 0.320 | |
| m × lv × ss | 40 | 0.015 | |

*Note.* Listed effects are statistically significant at $\alpha = 0.01$ and are sized $\eta^2 > 0.01$.

(a) normal items, (b) normal and bimodal items, and (c) normal, left-skewed, and right-skewed items. Scales consisting of right-skewed and normal items result in an overestimation of the LV skewness, and, analogously, scales consisting of left-skewed and normal items result in an underestimation of the LV skewness for FA-lin and IRT-mok.

For the right skew-normal LV, the results are generally worse than for the normal LV conditions. Here too, we observe patterns similar to those found for the normal LV conditions, contrasting FA-poly and IRT-grm to FA-lin and IRT-mok on the other hand. This contrast is even more distinct in the conditions involving a skew-normal LV. From the boxplots it can be observed that the distributions of all results are left-skewed, indicating a rather extreme underestimation of the skewness in a number of replications. FA-poly and IRT-grm results are best for the scales with normal and bimodal items. Skewness is underestimated in the large majority of replications, and more so for FA-poly than for IRT-grm.

For FA-lin and IRT-mok, skewness is also underestimated in most cases. However, when the scale consists of right-skewed and normal items, skewness is overestimated, though not as much as it is in the normal LV condition. Remarkably, the scale consisting of normal, right-skewed, and left-skewed items produces the best results for FA-lin. IRT-mok results deviate most from FA-lin in this condition.

A better approximation of the LV distribution by FA-poly and IRT-grm than by FA-lin and IRT-mok is presumably caused by the fact that the former two models take into account the threshold estimates. As these estimates were found to be better for IRT-grm than for FA-poly in case of a right skew-normal LV, it follows that IRT-grm LV estimates more closely resemble the true LV scores than FA-poly LV estimates do.

Furthermore, results are relatively better for symmetric item distributions and a balanced scale of as many right-skewed items as left-skewed items, in case of a normal LV.

### Association Between Sampled and Estimated LV Scores

The last column of Table 6.10 presents the ANOVA results with Kendall's $\tau_a$ between the true and estimated LV scores as the response variable. Judging from the $\eta^2$ values we can conclude that there are three important main effects: LV distribution ($\eta^2 = 0.262$), scale shape ($\eta^2 = 0.200$), and model ($\eta^2 = 0.095$) to a lesser extent.

The associations between the LV true scores and their estimates are illustrated by the scatterplots of Figures 6.24 and 6.25 for the normal and skew-normal LV conditions, respectively, for $n = 200$. Plots for the medium sample size condition can be found in Appendix E.4, Figures E.4 and E.5. To focus on the variations between design cells and cancel out the average variation between replications, the deviation scores $\theta_{sr} - \bar{\theta}_r$ and $\hat{\theta}_{sr} - \bar{\hat{\theta}}_r$, where $\theta_{sr}$ is the LV score of respondent $s$ in repetition $r$, of the true and estimated scores are depicted on the horizontal and vertical axis, respectively. The rows of the figures represent cells of the design and the columns the estimation models, which is indicated in the right lower corner of the subfigures. Each subfigure contains $R = 1000$ superimposed scatterplots. The white lines depict

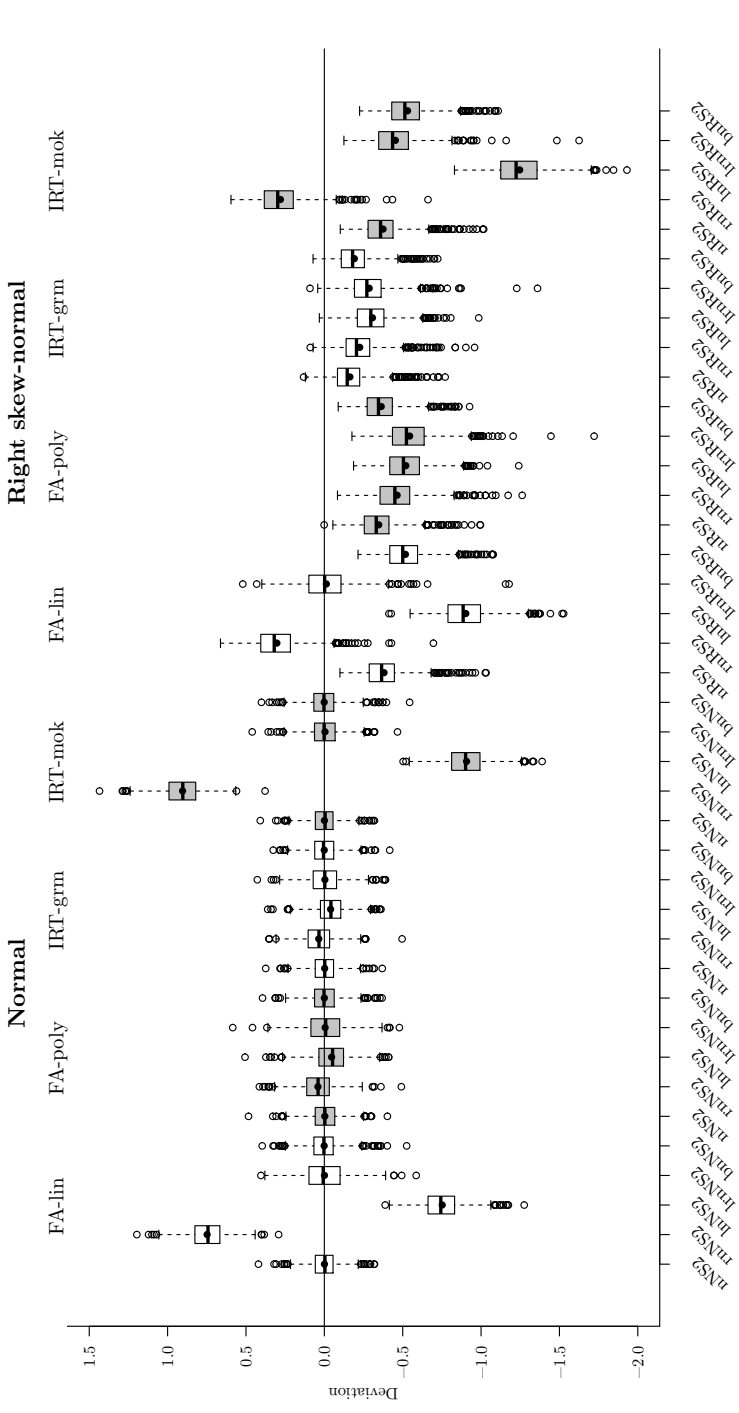



*Figure 6.22.* Deviation of skewness of estimated LV scores from skewness of sampled true LV scores for normal and right skew-normal LV conditions. $n = 200$; $R = 1000$.

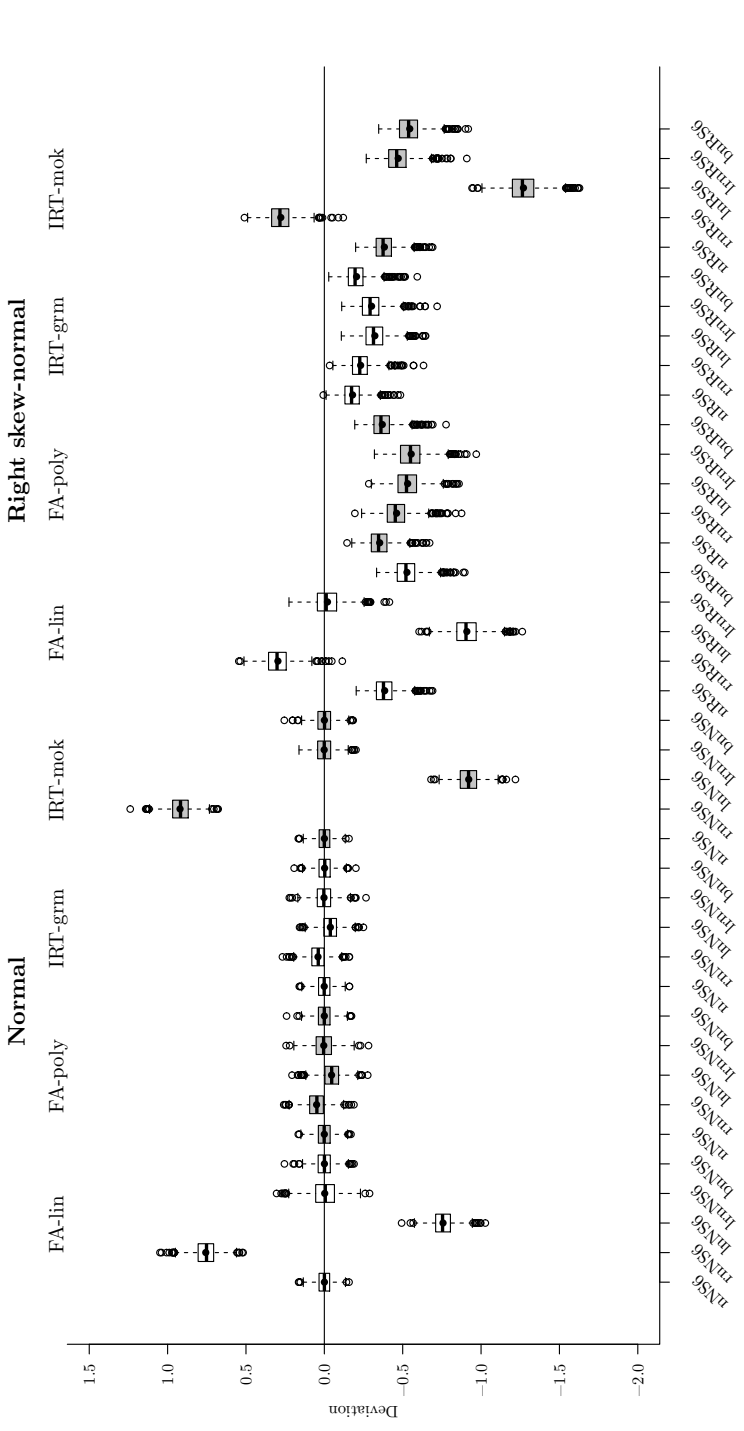



*Figure 6.23.* Deviation of skewness of estimated LV scores from skewness of sampled true LV scores for normal and right skew-normal LV conditions. $n = 600$; $R = 1000$.

the identity association. As IRT-mok LV scores are not on a latent scale, but on the integer scale of sum scores, the scatterplot of these estimates and the population LV scores does not approximate the white line, whereas the other scatterplots are centered around it. The minimum, median, and maximum Kendall's $\tau_a$ are given in the upper left corner inset of each subfigure.

We first note that, in general, the association between the estimated and population LV scores is quite strong with Kendall's $\tau_a$ values ranging between 0.73 and 0.89.

When the LV distribution is normal, FA-poly and IRT-grm estimates are closely and linearly related to their true counterparts, regardless of the item distributions. The scores in the tails of the distributions are estimated relatively well then, compared to FA-lin and IRT-mok. FA-lin estimates show an S-shaped relation to the true scores in case the scale consists of normal and right-skewed, left-skewed, or bimodal items. FA-lin LV estimates for scales including only normal items or both right- and left-skewed items are unaffected. For IRT-mok the same approximately holds as for FA-lin, except for the bimodal scale, which more closely resembles the results of the normal-scale than of the right- or left-skewed item scales.

When the LV distribution is skew-normal, the association between the estimated and true LV scores is slightly weaker than in the normal LV conditions. A curvilinear association is then apparent for each model and each condition. The flattened top visible in the IRT-mok plots illustrates the effect that for the right skew-normal LV it is not possible to distinguish between respondents in the upper tail of the distribution (say, $\theta_s > 8$) based on their item responses. For the parametric models, the top true LV scores are not estimated accurately either, judging from the cloud of points in the upper right corner of the graphs.

**Summary**

The right skew-normal LV causes estimation problems in the right tail of the distribution, as there is little information available from these respondents. Both the estimation of the shape of the LV distribution and the ordering of respondents are affected here. Nonnormal item distributions do not seem to influence FA-poly and IRT-grm LV score estimation much, as long as the LV distribution is normal, which is in accordance with Expectations 3t and 4t. FA-lin and IRT-mok LV score distributions are not recovered accurately in those cases, supporting Expectations 1h, 2k, and 5b, whereas the ordering of respondents is not affected much.

For each condition, IRT-mok LV score estimates, i.e., simple sum scores, are associated strongest with the population values. We can therefore conclude that, if a scale consists of items that are equally strongly related to the LV (i.e., equal item loadings), weighting by estimated loading and/or threshold values does not improve the estimated ordering of respondents on the LV, compared to simply taking the sum scores. In case of a skew-normal LV, the *ordering* of respondents using simple sum scores is even better than using weighted sum scores, when item loadings are equal. Compared to the other models, the *shape* of the LV distribution is best recovered

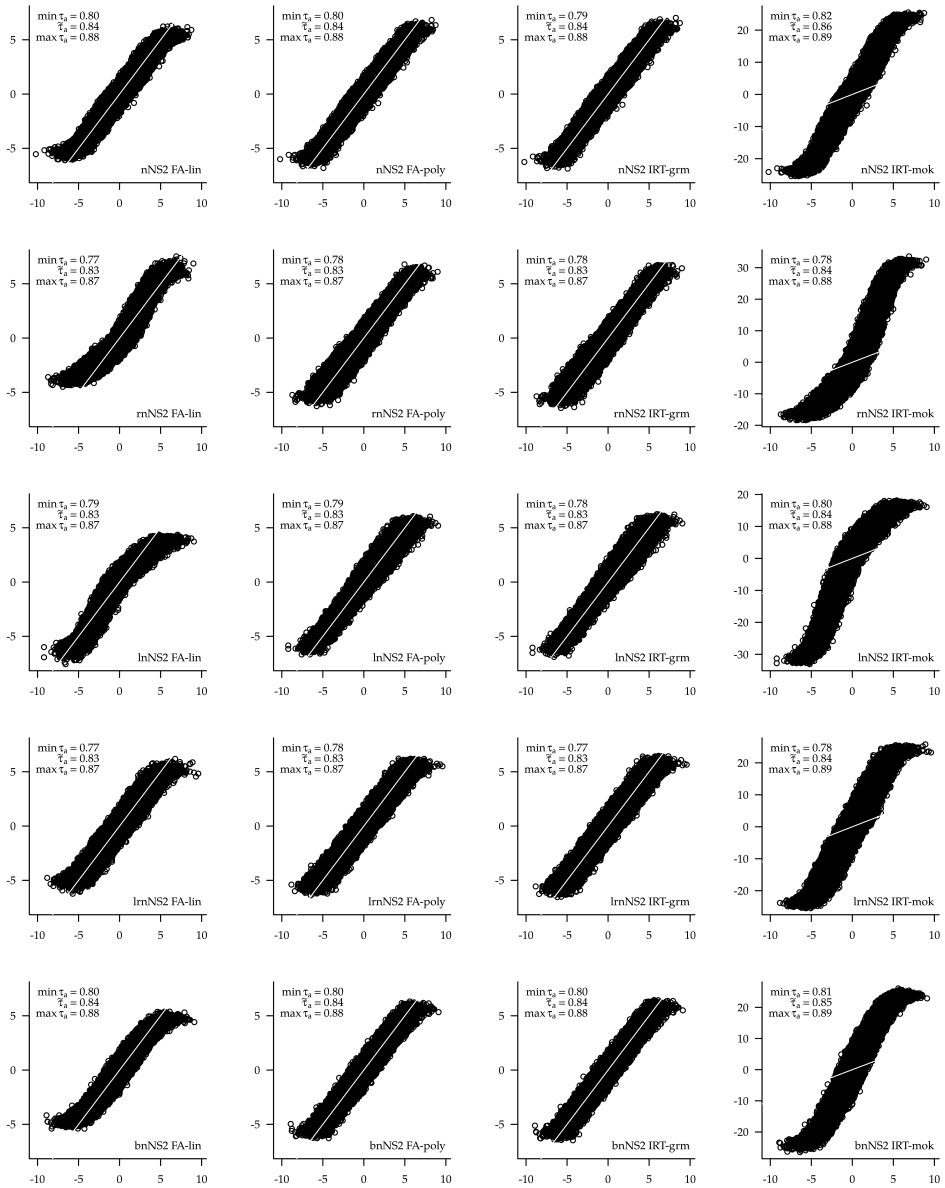



*Figure 6.24.* Scatterplots of LV population $\theta_{sr} - \bar{\theta}_r$ ($x$-axis) and estimated $\hat{\theta}_{sr} - \bar{\hat{\theta}}_r$ ($y$-axis) deviation scores for FA-lin, FA-poly, IRT-grm, and IRT-mok in Cells nNS2, rnNS2, lnNS2, lrnNS2, and bnNS2 for each replication. The minimum, median, and maximum Kendall's $\tau_a$ over replications are given in the inset of each plot as an indication of association. $n = 200$; $R = 1000$.

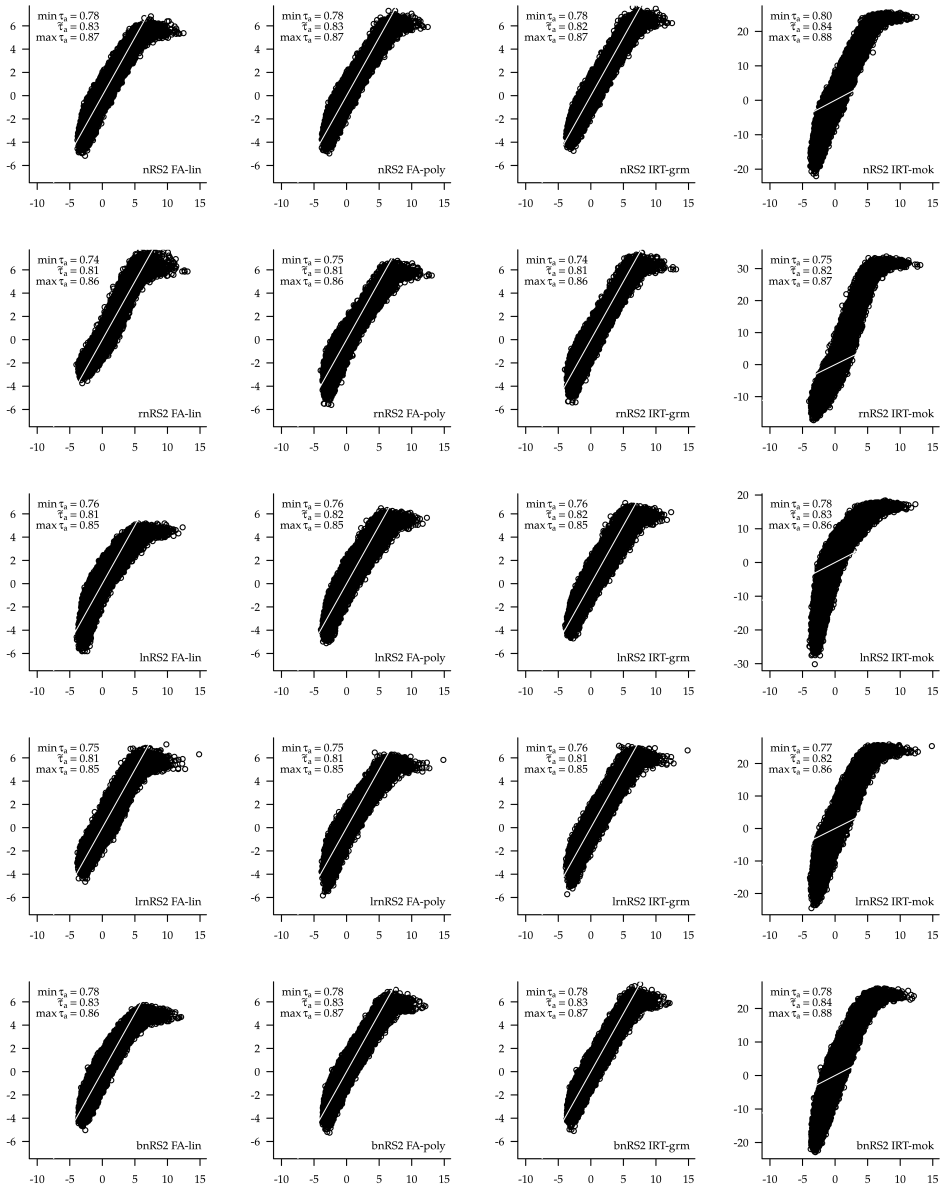



*Figure 6.25.* Scatterplots of LV population $\theta_{sr} - \bar{\theta}_r$ ($x$-axis) and estimated $\hat{\theta}_{sr} - \bar{\hat{\theta}}_r$ ($y$-axis) deviation scores for FA-lin, FA-poly, IRT-grm, and IRT-mok in Cells nRS2, rnRS2, lnRS2, lrnRS2, and bnRS2 for each replication. The minimum, median, and maximum Kendall's $\tau_a$ over replications are given in the inset of each plot as an indication of association. $n = 200$; $R = 1000$.

by IRT-grm for the skew-normal LV distribution, except when the scale consists of left-skewed, right-skewed, and normal items, in which case the distribution of FA-lin LV score estimates approximates the sampled LV score distribution best.

## 6.3 Discussion

In this chapter, we presented the results of applying the four scaling models to a number of data configurations, representing violations of distributional conditions. We compared FA of the sample covariance matrix (FA-lin), FA of the estimated polychoric correlation matrix (FA-poly), the graded response IRT model (IRT-grm), and the nonparametric Mokken IRT model (IRT-mok) on a number of performance criteria regarding estimators of parameters, corresponding standard errors, model fit, and latent variable (LV) scores. For parameter and standard error estimators, we examined both the accuracy, as reflected by the absence of bias of the estimator, and the precision, as indicated by the low dispersion of the estimates.

Data were generated under the FA-poly model, with either a normal or a right skew-normal LV distribution, and with variously shaped item distributions, which were manipulated independently of the LV distribution. The main purpose of the study was to determine the robustness of the parametric models against these violations of model assumptions, and to compare the parametric models' performance to the behavior of the nonparametric model in conditions of nonnormality.

In Table 6.11 a summary of most of the results for the parametric models is presented referring to the expectations brought forth in Chapter 4. The left and right parts of the table represent the normal and the right skew-normal LV conditions, respectively. The upper and lower panels represent the normal and skewed item conditions, respectively. Therefore, the upper left quadrant of the table represents the normal conditions discussed in Chapter 5.

From the left part of the table it is apparent that in case of a normal LV distribution FA-poly and IRT-grm perform well with regard to every performance variable included in our design, regardless of the item distribution. FA-lin is clearly outperformed.

From the right part of the table we can infer that the performance of all models deteriorates as a result of a skew-normal LV, most notably when combined with skewed item variables (the lower right quadrant). We also observe that IRT-grm performs best in such conditions. In the following, we will elaborate on these findings.

### Parameters

Compared to the other models under investigation, we generally found loading parameters estimated by FA-lin to be the most biased, which supports Expectation 1b. FA-lin parameter bias was within the acceptable range of 5% deviation only for the right-skewed item loading on the right skew-normal LV, presumably because the thresholds were spaced exactly equally in this condition. In that condition, relatively good performance was expected (Expectation 2c), but parameter estimation was even more

*Table 6.11.* Summary of results referring to the expectations presented in Section 4.3 (p. 91). Results deviating from the expectations are printed larger and in boldface.

| | LV | Normal distribution | | | | Right skew-normal distribution | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | $\hat{\omega}$ | $\hat{se}(\hat{\omega})$ | Model | LV score | $\hat{\omega}$ | $\hat{se}(\hat{\omega})$ | Model | LV score |
| Item | Model | Bias | Bias | Fit | Bias | Bias | Bias | Fit | Bias |
| Normal | FA-lin | − | **√**[a] | √ | **√/−**[b] | − | **√**[a] | √ | − |
| | | 2a | **2e** | 2f/2g | **2k** | 2d | **2e** | 2f | 2k |
| | FA-poly | √ | √ | √ | √ | **√**[g]/√/− | √ | √ | √ |
| | | 3a/3b | 3l/3n | 3q/3s | 3t | **3e**/3h | 3l | 3r/3s | 3t |
| | IRT-grm | √/ | √ | √ | √ | **√**[d] | + | √ | √ |
| | | 4a/4b/4c/4d | 4l/4n | 4q/4s | 4t | **4i** | 4p | 4r/4s | 4t |
| Skewed | FA-lin | − | **−/√**[e] | −/√ | −/− | −/**√**[f] | − | −/**−**[g]/√ | − |
| | | 2b | **2e** | 2h/2i/2j | 2k | **2c** | 2e | 2h/**2i**/2j | 2k |
| | FA-poly | √ | √ | √ | √ | +/−/**±**[h]/**±**[i] | − | √ | √ |
| | | 3c/3d | 3m/3o | 3r/3s | 3t | 3f/3g/**3i**/**3j** | 3p | 3r/3s | 3t |
| | IRT-grm | √ | √ | √ | √ | **√**[j]/−/**√**[k]/**√**[l] | √ | √ | √ |
| | | 4e/4f | 4m/4o | 4r/4s | 4t | **4g**/4h/**4j**/**4k** | 4p | 4r/4s | 4t |

*Note.* √ indicates good performance; +, −, and ± indicate positive bias, negative bias, and a combination of both, respectively.
[a] Loading standard errors expected −. [b] LV scores expected ±, but only in tails.
[c] Loading parameters expected ±. [d] Outer thresholds expected −. [e] Loading standard errors expected −, but not for mixed-skewed items.
[f] Loading parameters expected the same as in normal conditions. [g] RMSEA expected √, but only mean value. [h] All threshold parameters expected +, but first and second − for right-skewed items. [i] All threshold parameters expected +, but second √ and third and fourth − for left-skewed items. [j] Loading parameters expected + for right-skewed items. [k] Threshold parameters expected + for right-skewed items.
[l] Threshold parameters expected + for left-skewed items.

accurate than in the normal condition then. This difference is probably due to the fact that thresholds were spaced exactly evenly in the condition of right-skewed items and a right skew-normal LV, whereas the spacing was only approximately even in the normal-normal condition. Parameter estimators were negatively biased up to −38%, which occurred for left-skewed items in a scale of normal, left-skewed, and right-skewed items loading on a right skew-normal LV.

FA-poly loading and threshold parameters were unbiased for every item distribution as long as the LV distribution was normal, supporting Expectations 3c and 3d. In case of a skew-normal LV, parameter estimation worsened considerably, resulting in biased loading and threshold parameter estimators for the skewed item distributions (7.3% for right-skewed and −12.8% for left-skewed items), which is in accordance with Expectations 3f and 3g. Discrimination parameter results display a pattern similar to the loading parameter results. Effects are enlarged, however, presumably because of the logit scale of these parameters.

IRT-grm loading and threshold parameter estimators appeared to be the most robust against the violations under investigation. Only for the left-skewed items loading on the right skew-normal LV, did we find an unacceptable negative loading parameter estimation bias of −6.5% at most, which is consistent with Expectation 4h. Threshold parameters of left-skewed items were found to be unbiased, although we expected them to be overestimated (Expectation 4k). For the right-skewed items both loading and threshold parameter estimators were unbiased, which is also better than expected (Expectations 4g and 4j).

### Standard Errors

In case of a normal LV distribution, FA-lin standard error estimators were negatively biased for right-skewed and left-skewed items, in the absence of oppositely skewed items. In the skew-normal LV conditions, all FA-lin standard error estimators were severely biased (up to 49%), with underestimation for right-skewed items and overestimation for left-skewed items. These findings support Expectation 2e and hopefully settle the inconclusiveness on the robustness of FA-lin standard errors found in the literature, as FA-lin loading standard errors were found to be overestimated (Babakus et al., 1987) as well as underestimated for skewed items loading on a normal or skew-normal LV (DiStefano, 2002; B. O. Muthén & Kaplan, 1985; Rhemtulla et al., 2012).

FA-poly loading standard error estimators were unbiased in case of a normal LV. Under a right skew-normal LV distribution, unacceptable negative bias was found for right-skewed items (−13.5% bias) and for normal items in scales that include right-skewed items (about −10% bias). These results for the right-skewed items loading on a right skew-normal LV support Expectation 3p. The underestimation of standard errors is in line with the underestimation encountered to a lesser extent in the normal conditions. As neither normal, bimodal, nor left-skewed items loading on a skew-normal LV were included in any previous research taking standard error

estimation into consideration, the standard error results for these item distributions are unprecedented.

IRT-grm loading standard error estimators were characterized by high accuracy but relatively low precision, as the standard error estimators were unbiased for every item distribution, supporting Expectations 4m and 4p, but also exhibited the largest variance of all estimation models.

Discrimination standard error estimation results resembled the results of the loading standard errors for both FA-poly and IRT-grm.

Threshold standard error estimators did not deviate substantially from the empirical standard deviations for either FA-poly or IRT-grm in any of the conditions, and are therefore considered robust. With regard to the normal LV conditions, these results are in accordance with Expectations 3o and 4o; for the skew-normal LV conditions, these results are unprecedented.

### Model Fit

Model fit was asserted using the $\chi^2_{YB}$, the RMSEA based on the $\chi^2_{YB}$, and the SRMR. Of all models, FA-lin average $\chi^2_{YB}$ values showed the largest deviation from the expected values. The deviation was largest for scales that contain skewed items, and increased with the number of skewed items included in the scale, supporting Expectation 2h. Although on average the RMSEA values did not exceed the criterion of 0.06, we found that the distribution of RMSEA estimates deviated considerably from the theoretical distribution when skewed items were involved. Consequently, the RMSEA as estimated by FA-lin is not suitable for such conditions. These results refine the conclusions of research by DiStefano (2002) who reported some robustness of FA-lin RMSEA against skewed items, based on the average RMSEA and 95%-confidence intervals for the population value only.

For FA-poly and IRT-grm, we found that the $\chi^2_{YB}$ statistic and the $\chi^2_{YB}$-based RMSEA are very useful for both FA-poly and IRT-grm modeling in normal as well as nonnormal distributional conditions, supporting Expectations 3r, 3s, 4r and 4s

The SRMR results were indicative of a good fit in all cases for FA-lin and FA-poly, consistent with Expectations 2j and 3s. For IRT-grm the average SRMR reached or exceeded the performance criterion of 0.08 only in conditions of scales including left-skewed items loading on a right skew-normal LV. In other conditions, IRT-grm also resulted in a larger SRMR than the FA models did. This seems somewhat contradictory to the equally good to better parameter estimation results for IRT-grm compared to FA-lin and FA-poly. However, the SRMR is based on a comparison of model-implied and sample covariances, thus taking only the bivariate sample item information into account, which is general in FA, but not in IRT modeling. As FA modeling is based on covariances, parameters are estimated to minimize the discrepancy between model-implied and sample covariances. IRT-grm modeling takes into account the full response patterns, and is therefore not directly aimed at minimizing that discrepancy. So, because IRT-grm parameter estimation is not aimed at recovering covariances, SRMR results are relatively bad compared to the parameter estimation results.

As the SRMR is generally not available for IRT-grm, it had not been investigated before. It is of interest to further investigate the performance of the SRMR in case of model-data *mis*fit, since such conditions were not included in our design.

### Nonparametric IRT-mok Model

The nonnormal LV and item distributions did not pose any estimation problems for IRT-mok. On the contrary, scales with heterogeneously shaped item distributions led to increased $H$ values that were estimated with less bias. The greatest potential problem for an IRT-mok analysis is a scale of items with equal item means. In those cases, item ordering is arbitrary and causes problems for the $H$ coefficient which is built upon an item order.

The small positive parameter estimation bias for the normal scale was also found for the scale of normal items loading on the skew-normal LV. In the other conditions, parameters were also overestimated, but to a much lesser extent.

Standard error estimators were found to be unbiased in all conditions, and are thus considered useful for the interpretation of $H$ coefficients regardless of the LV or item distributions involved. Therefore, we conclude that IRT-mok is very well suited for data from nonnormal LV and item distributions, supporting Expectations 1f and 5a.

### Latent Variable Scores

LV score estimation was evaluated by considering the skewness of the estimates in comparison to the population skewness, and by examining scatterplots of the true and estimated LV scores and the corresponding Kendall's $\tau_a$ measure of association.

We found nonnormal item distributions not to influence LV score estimation considerably, as long as the LV distribution was normal. In these conditions, FA-poly and IRT-grm estimates were strongly and linearly related to the population values, supporting Expectations 3t and 4t. Although the ordering of respondents was not affected much, LV scores in the tails of the distributions were estimated more accurately by FA-poly and IRT-grm than by FA-lin and IRT-mok, which is in accordance with Expectation 1h.

The right skew-normal LV caused estimation problems in the right tail of the distribution, both in the distributional shape of the estimates and the ordering of respondents. For each condition, IRT-mok LV score estimates were related most closely to the population values.

We conclude that, if a scale consists of items loading equally on the LV, weighting by estimated loading and/or threshold values does not improve the ordering of respondents on the LV over simply taking the sum scores. In case of a skew-normal LV, simple sum scores even result in a better ordering compared to weighted sum scores, when item loadings are equal.

However, to assess the shape of the LV distribution, weighting by loading and threshold estimates is considered essential. Moreover, the results for the mixed-strength scale from the previous chapter indicated a clear benefit of weighting, as

FA-lin, FA-poly, and IRT-grm LV scores were superior to IRT-mok LV scores when item loadings differed among items within a scale. This finding calls for some further research involving mixed-strength scales of nonnormal LV and item distributions.

#### Sample Size

Overall, results were better for the medium ($n = 600$) than for the small ($n = 200$) sample size, which is in accordance with Expectation 1i. Parameter and standard error accuracy were not significantly improved going from the small to the medium sample size, except for the step-difficulty and item scalability parameters. Lack of precision was, in general, an issue for the small sample size, impairing the reliability — and hence the usefulness — of a single parameter or standard error estimate. For the medium sample size, precision much improved for all parameter and standard error estimators. These findings apply both to the normal and the nonnormal conditions.

#### Remarkable Findings

Bimodal item distributions have not been subject to much, if any, Monte Carlo research yet. In our implementation with the set of item response category proportions of $\{0.10, 0.35, 0.10, 0.35, 0.10\}$, we found that results for bimodal items were very similar to those for normal items. Presumably, this can be accounted to the distributional symmetry of the bimodal items, which is a preferable shape for model estimation.

Another remarkable finding is the difference in best performing scale configuration between the parametric models and IRT-mok: For IRT-mok the all-normal item scale is recovered worst (due to the location of the items), while for the parametric models it is (one of) the configuration(s) recovered best.

## 6.4 Recommendations

Based on the findings discussed in the previous section, we provide the following recommendations.

- For the sample size of 200, all models produced relatively unprecise estimators. None of the models stood out in performance then. Even IRT-mok's $H$ value estimation was much more precise for the sample size of 600. As the reliability and usefulness of a single parameter or standard error estimate is questionable, employing a sample size as small $n = 200$ is not recommended.

- When items are categorical (rather than continuous), FA-lin should not be employed. Parameter estimators are negatively biased, unless thresholds are equally spaced — an assumption that is quite impossible to check. Furthermore, in case of item and/or LV skewness, standard errors are negatively biased and model fit cannot reliably be estimated by means of the RMSEA.

- FA-poly is the right choice for any kind of item distribution, as long as the LV distribution is normal. This requires some knowledge of the population of interest with regard to the trait or ability one intends to measure using the scale.

- In case of a skew-normal LV, our findings show that IRT-grm is best employed for scale analysis, as it was least affected by nonnormality of the LV distribution. Only in the, perhaps unusual, case of left-skewed items loading on a right skew-normal LV, did IRT-grm results show some parameter and standard error estimation bias.

### 6.4.1 Inferring the LV Distribution

From these recommendations it is clear that much depends on the population LV distribution, which we unfortunately do not know in practice. Given our results, however, one can make some cautious inferences about the true LV distribution, based on sum score distributions and estimated LV score distributions, which can both be observed. Such inferences are useful in choosing the model whose results are likely the most accurate and reliable. Cautiousness is required here, because our design did not cover all possible combinations of item and LV distributions.

In the following, we shall explain how inferences about the shape of the population LV distribution can be made based on the shape of the distribution of sum scores and the shape of the LV score distribution as estimated by FA-poly and IRT-grm. Our reasoning is based on the LV skewness estimated by each of our four estimation models, as displayed in Figure 6.26. To facilitate within-cell and between-model comparisons, the skewness estimates are ordered by design cell and estimation model there.

Unweighted sum scores are taken as the IRT-mok LV estimates, as is commonly done in practice. Theoretically, however, IRT-mok merely provides an ordering of respondents, leaving the shape of the LV distribution undefined or unidentified. Because in practice one might want to make inferences about the LV distribution, we include IRT-mok LV scores in the comparison of the estimated LV distribution to its true counterpart, as applied to all estimation models.

From Figure 6.26 it is clear that when the IRT-mok LV score distribution has zero skewness, the true LV distribution is probably also normal, except in case of left-skewed items loading on a right skew-normal LV (and, presumably, vice versa). Therefore, if the sum score distribution is normal, the LV distribution can probably be well represented by a normal LV. After applying FA-poly or IRT-grm, one can assess the shape of the model-based LV score distribution. If either FA-poly or IRT-grm LV score estimates are normally distributed, in addition to the sum scores, it is very likely that the true LV distribution is normal, since the FA-poly and IRT-grm LV skewness values are only zero in case of a normal LV (see Figure 6.26). In case the FA-poly or IRT-grm LV distribution is *not* normal, given the normal sum scores, the true LV distribution presumably also deviates from normality.

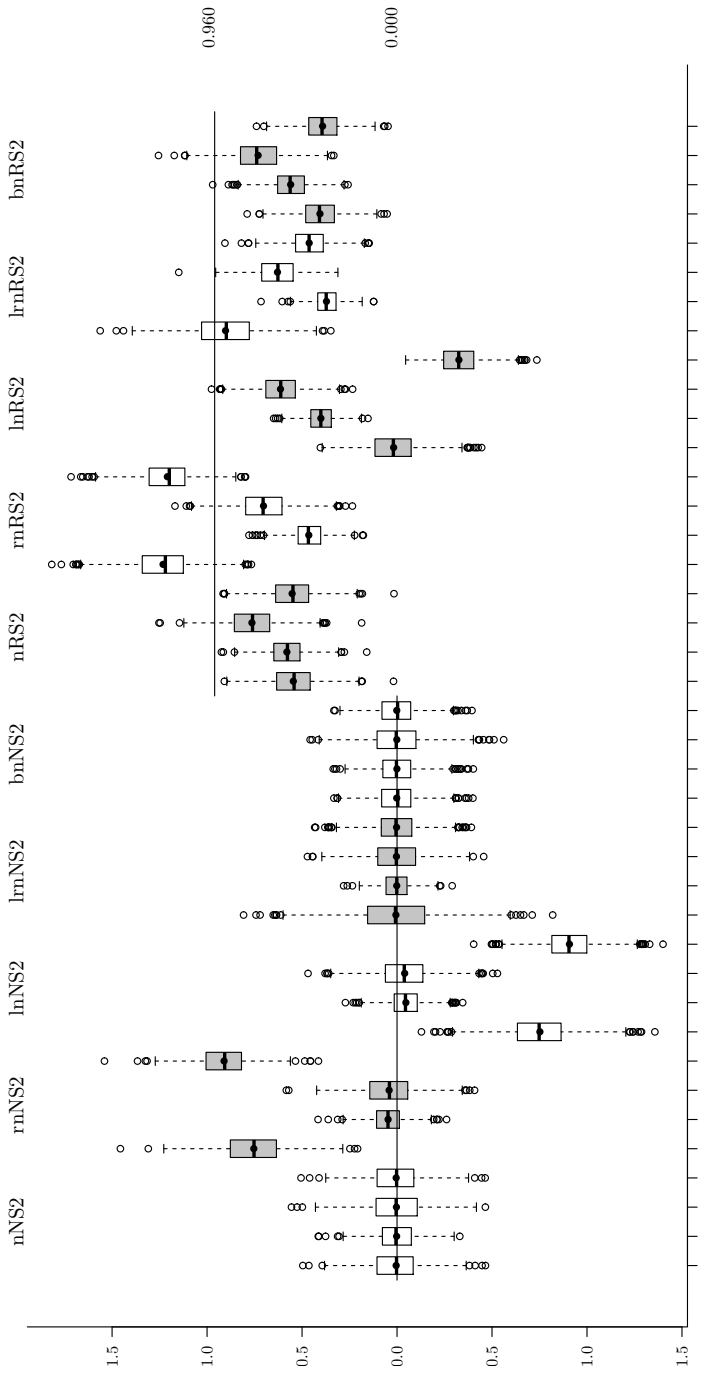



*Figure 6.26.* Skewness of estimated LV scores for normal and right skew-normal LV conditions. Population skewness is indicated in the right margin. $n = 200$; $R = 1000$.

When the sum score distribution is not normal, it is advised to apply the IRT-grm model — preferably supplemented by the FA-poly model — and assess its estimated LV distribution. If the IRT-grm estimated LV distribution is normal, then the true LV is probably normal. If the IRT-grm estimated LV distribution is skewed in the same direction as the sum score distribution, the true LV is probably skewed, most likely more severely than the estimated LV distribution (cf. Cells nRS2, lrnS2, and bnRS2 in Figure 6.26). In case of a congruent direction of item skewness and LV skewness, sum score skewness will be larger than IRT-grm-estimated skewness, and the population LV skewness is expected to be somewhere in the middle (cf. Cell rnRS2 in Figure 6.26). If the IRT-grm estimated LV distribution is skewed in the opposite direction from the sum score distribution skewness, the true LV is probably skewed in the direction indicated by IRT-grm (cf. Cell lnRS2 in Figure 6.26). Under conditions of LV skewness, the skewness values of the LV distribution as estimated by FA-poly and IRT-grm are more divergent than when the true LV is normal, which is why it is preferable to apply both FA-poly and IRT-grm.

In case FA-poly or IRT-grm are indicative of a skew-normal LV distribution, most likely the true LV distribution is even more skewed, as this is the case in all skew-normal conditions of our design (see Figure 6.26). In addition, the underestimation of skewness is smallest when items are either normal or bimodal (symmetrical), and larger for skewed items, regardless of whether the direction of item skewness equals that of the LV skewness.

These types of inferences can be used as a guide when applying scale analysis in practice, as will be demonstrated in the next chapter.

# Chapter 7

# Applications of FA and IRT

## 7.1 Introduction

In this chapter, we return to the *practice* of scale analysis, by applying the scaling models under investigation, factor analysis of the sample covariance matrix (FA-lin), factor analysis of the estimated polychoric correlation matrix (FA-poly), the graded response item response theory model (IRT-grm), and the nonparametric Mokken item response theory model (IRT-mok), to empirical data. Using these empirical illustrations, we demonstrate how to apply the findings from our simulation study presented in the previous chapters.

We selected three scales designed to measure various psychological and sociological latent variables (LVs). The first scale under investigation is the Dresden Body Image Questionnaire (DBIQ; Pöhlmann, Thiel, & Joraschky, 2008; Scheffers, Van Duijn, Bosscher, Wiersma, & Van Busschbach, 2013), presented to a community sample of size $n = 761$, comparable to the medium sample size of our simulation study. The DBIQ consists of 35 items loading on five LVs. We performed five unidimensional analyses and a multidimensional analysis.

The second scale we analyzed is the Revised Anticipated Sexual Jealousy Scale (RASJS; Buunk, 1982, 1997), using data of a community sample of size $n = 1366$. The RASJS consists of three subscales of five items each, reflecting different aspects of sexual jealousy. We use these data to demonstrate the behavior of the scaling models when a theoretically multidimensional scale is analyzed using a simplified model, regarding all subscales as one scale representing the LV one level higher.

The third application concerns the short version of the Involvement in Neighbourhood Community Scale (INCS; Frieling, 2008), presented to a community sample of size $n = 255$, comparable to the small sample size of our simulation study. The INCS contains seven items.

## 7.2 Setup of the Analyses

All data sets were analyzed by means of FA-lin, FA-poly, IRT-grm, and IRT-mok. The results of the parametric models are presented simultaneously, followed by the nonparametric IRT-mok results.

The parametric models were estimated using MPLUS (L. K. Muthén & Muthén, 1998–2010). IRT-mok was applied using the R package `mokken` (Van der Ark, 2011).

We begin each results section by examining the LV scores as estimated by each of the applied scaling models to decide which estimation model is most appropriate and to facilitate the interpretation of the additional results.

For each subscale, we first examine the shape of the sum score distribution. If it is normal, it is to be expected that a normally distributed LV is suitable for representing the data. Second, the shapes of LV distributions as estimated by the parametric models are inspected. From our simulation study we know that the skewness of the LV distribution is best recovered by IRT-grm, so we generally expect IRT-grm to produce the best LV estimates. We should note that in making these inferences we assume the model structure to be correctly specified, as was the case in the simulation study. In the selected empirical examples this assumption is rather plausible, because they all concern scales that have already been analyzed employing item selection.

From the inferred information on the LV distribution, expectations on the pattern of differences between the parameter and standard error estimates of the three parametric models follow, based on the Monte Carlo study results. These hypothesized patterns are checked and similarities and differences between model results are interpreted.

## 7.3 Dresden Body Image Questionnaire

Recently, a Dutch version of the Dresden Body Image Questionnaire (DBIQ; Dresdner Körperbildfragenbogen; Pöhlmann et al., 2008) was proposed by Scheffers et al. (2013). In two non-clinical convenience samples they investigated the psychometric properties of the translated 35-item scale, using confirmatory linear FA. The DBIQ is a five-point Likert scale consisting of 35 positively or negatively worded items to which respondents have to rate their level of agreement. The DBIQ consists of five subscales, *Vitality* (8 items), *Body acceptance* (8 items), *Sexual fulfillment* (6 items), *Self-aggrandizement* (7 items), and *Physical contact* (6 items).

The purpose of the scale is to develop a valid and reliable measure for the multidimensional construct *body image* to be used in non-clinical and clinical populations, as *body image* is related to psychosocial functioning in general and to a wide range of psychiatric disorders.

Here, the data of the larger sample ($n = 761$, 433 women, 326 men, 2 unknown gender) of the two samples discussed in Scheffers et al. (2013) are reanalyzed to investigate and compare the performance of the FA and IRT models under investigation, and to relate these results to the findings from our simulation study.

Although Scheffers et al. (2013) originally applied a multidimensional analysis, taking into account the dependencies of the five aspects of *body image* represented by the subscales, for our illustrative purposes we applied five unidimensional analyses in correspondence with our simulation study. To obtain a sense of the differences and similarities of a unidimensional and multidimensional approach, we complement our analyses with a multidimensional application of the parametric models, and briefly present some of the latter results.

As deleting the missing values listwise would result in disregarding 76 of the 761 cases, we chose to apply the parametric analyses on the complete data set, letting MPLUS take care of the missing values using full-information maximum likelihood (ML). In the `mokken` package used for IRT-mok, such an option is not available, so we resorted to listwise deletion for the nonparametric analysis.

We start with a description of the sample data. Subsequently, we present the results of applying the four scaling models. Finally, the DBIQ results are briefly discussed.

### 7.3.1 Descriptive Statistics

The items of the DBIQ and their means and standard deviations are given in Table 7.1, where the items are ordered by subscale. In the actual questionnaire, however, the items of the various subscales are blended. Some items were worded such that a low score represents a high level of the LV. Such items were coded reversely to facilitate the analysis and interpretation of results. For example, the response *"Fully"* to the first item of the *Vitality* subscale *"I often feel physically run down"* represents a low level of *Vitality* and is therefore coded as 0 rather than 4. The coding of items preceded all additional processing of the data. So, reversely coded items are plotted as such.

Graphs of the item and subscale distributions, and of the total score distribution can be found in Figures 7.1 to 7.6. In the inset of the item plots, the skewness ($\varsigma$), excess kurtosis ($\kappa$), and number of missing values (NA: not available) are given. The sum score plots contain the empirical mean and standard deviation as well.

The sum score distribution of the first subscale *Vitality* is slightly left-skewed ($\varsigma = -0.34$). The items are also all left-skewed. Item skewness ranges between $-0.44$ and $-0.84$. Excess kurtosis varies between $-0.06$ and 1.15.

The sum score distribution of the second subscale *Body acceptance* is left-skewed ($\varsigma = -0.75$). Item skewness ranges between $-1.06$ and $-0.14$, and excess kurtosis is between $-0.89$ and 0.81.

The sum score distribution of the third subscale *Sexual fulfillment* is left-skewed ($\varsigma = -0.87$). The items that make up that scale are all left-skewed with skewness ranging from $-0.98$ to $-0.31$, and excess kurtosis between $-0.27$ and 1.33.

The fourth subscale *Self-aggrandizement* has a normal sum score distribution ($\varsigma = -0.07$). Its items have a variety of shapes with skewness ranging from $-0.68$ to 0.36, and excess kurtosis between $-0.53$ and 0.73.

*Table 7.1.* Dresden Body Image Questionnaire items ordered by subscale.

| Abbreviation[a] | Item Description | Mean[b] | SD |
|---|---|---|---|
| v1.run.down | I often feel physically run down. (R)[c] | 2.95 | 0.84 |
| v2.motivation | I lack energy and motivation. (R) | 2.90 | 0.92 |
| v3.exhaust | I often feel physically exhausted. (R) | 2.89 | 0.80 |
| v4.fit | I am physically fit. | 2.89 | 0.83 |
| v5.energy | I have lots of energy. | 2.79 | 0.81 |
| v6.ph.cond | I am in good physical condition. | 2.90 | 0.79 |
| v7.ph.limits | I quickly reach my physical limits. (R) | 2.66 | 0.92 |
| v8.ph.strong | I am physically strong and resilient. | 2.94 | 0.83 |
| ba1.happy | There are lots of situations in which I feel happy about my body. | 2.82 | 0.80 |
| ba2.like | I like my body. | 2.56 | 0.85 |
| ba3.clothing | I choose clothing that hides the shape of my body. (R) | 2.88 | 0.98 |
| ba4.uncomf | I often feel uncomfortable about my body. (R) | 3.06 | 0.89 |
| ba5.wish.diff | I wish I had a different body. (R) | 3.15 | 1.00 |
| ba6.satisfied | I am satisfied with my appearance. | 2.76 | 0.77 |
| ba7.change | If I could change something about my body, I would do it. (R) | 2.59 | 1.19 |
| ba8.show | I like showing my body. | 1.83 | 0.91 |
| s1.intense | I experience intense and pleasurable feelings during sex. | 2.91 | 0.90 |
| s2.satisfied | I am very satisfied with my sexual experiences. | 2.80 | 0.93 |
| s3.important | I think sex is an important part of life. | 2.46 | 0.97 |
| s4.inhibit | I am able to lay aside my inhibitions in sexual situations. | 2.67 | 1.00 |
| s5.enjoy | I am able to enjoy my sexuality. | 2.92 | 0.89 |
| s6.satisfying | My sexual experiences are satisfying. | 2.77 | 0.94 |
| sa1.graceful | I move gracefully. | 2.26 | 0.90 |
| sa2.attractive | Other people find me attractive. | 2.43 | 0.69 |
| sa3.pleasant | I find it pleasant and exhilarating when someone looks at me attentively. | 2.62 | 0.89 |
| sa4.valued | I feel more valued when someone pays attention to my body. | 2.67 | 0.84 |
| sa5.expressive | My body is expressive. | 2.26 | 0.82 |
| sa6.use.body | I use my body to attract attention. | 1.25 | 0.93 |
| sa7.centre | I like to be the centre of attention. | 1.65 | 1.00 |
| ph1.close | Physical contact is important for me to express closeness. | 2.84 | 0.83 |
| ph2.search | I look for physical intimacy and affection. | 2.47 | 0.85 |
| ph3.touch | I do not like people touching me. (R) | 3.05 | 0.91 |
| ph4.hug | I like it when people put their arms around me. | 2.93 | 0.83 |
| ph5.avoid | I consciously avoid touching other people. (R) | 3.15 | 0.87 |
| ph6.few.people | I only allow a few people to touch me. (R) | 2.33 | 1.06 |

*Note.* The item descriptions were taken from Scheffers et al. (2013). Items were rated on a five-point Likert scale ranging from *"Not at all"* to *"Fully"*, represented by 0 and 4, respectively, except for the reversely coded items, for which 4 and 0 were used, respectively.

[a] v: *Vitality*, ba: *Body acceptance*, s: *Sexual fulfillment*, sa: *Self-aggrandizement*, ph: *Physical contact*. [b] Item means and standard deviations were computed from the sample of size $n = 761$ with itemwise removal of missing values.

[c] R: Reversely coded item. Recoding was performed before the mean was computed.

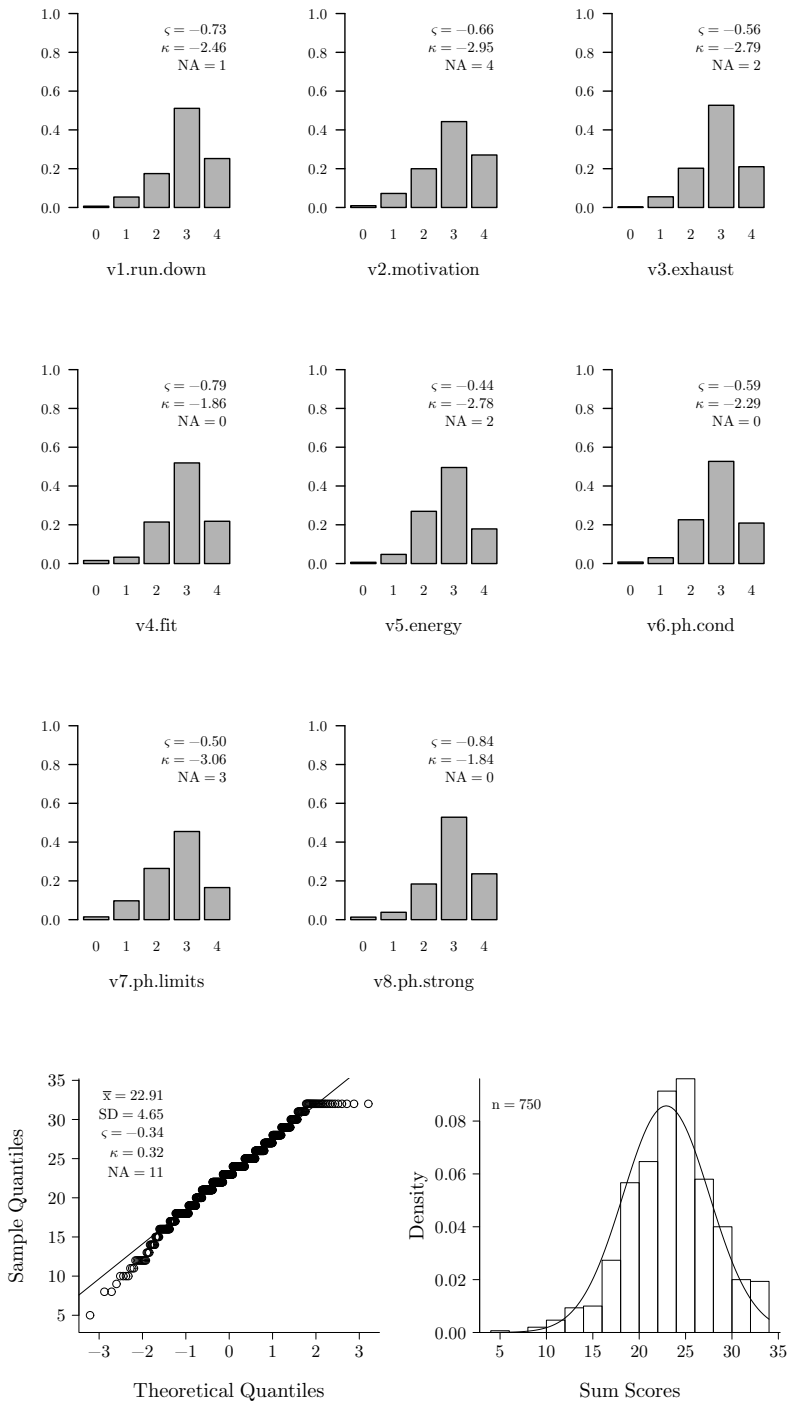



*Figure 7.1.* Item barplots and sum score Q-Q plot and density plot for DBIQ-*Vitality*. $n = 761$.

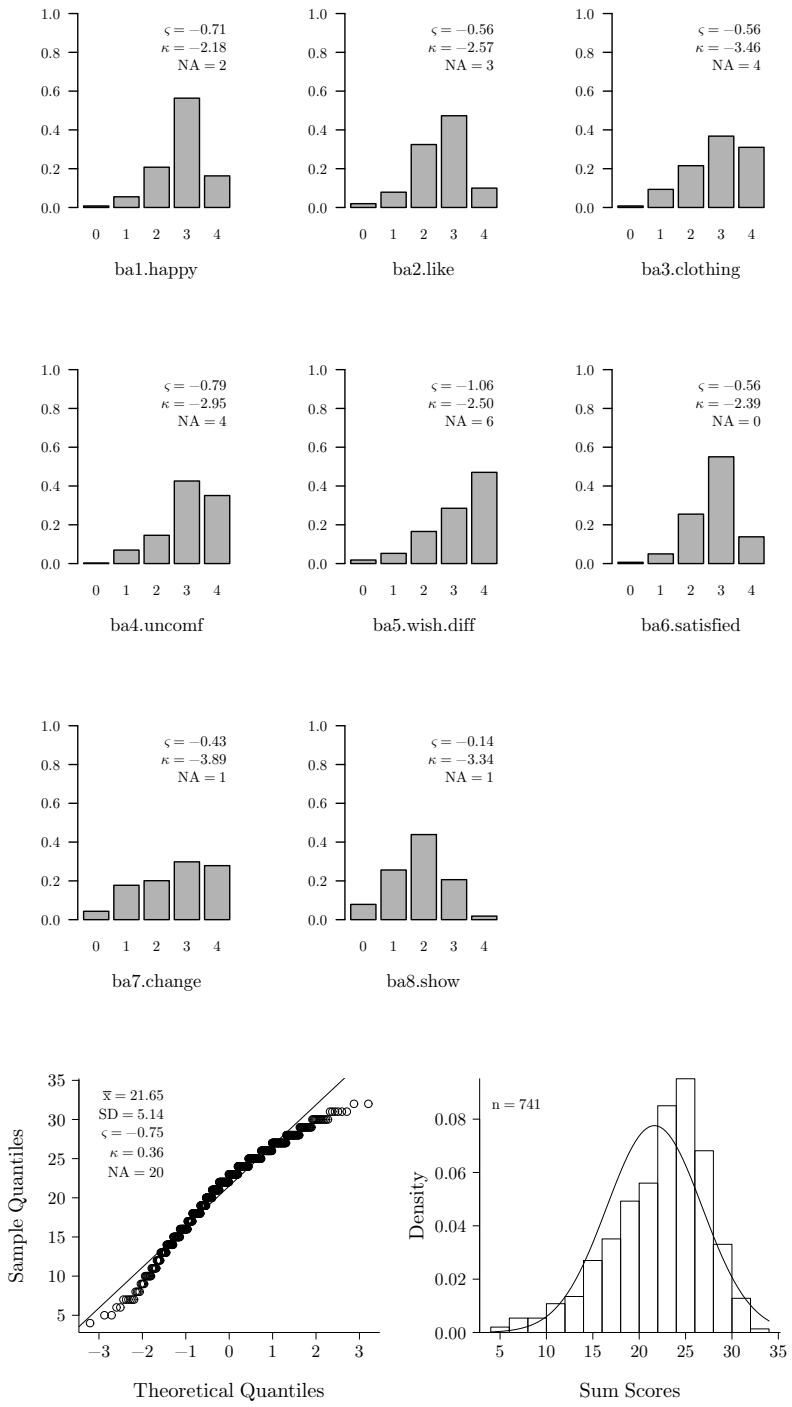



*Figure 7.2.* Item barplots and sum score Q-Q plot and density plot for DBIQ-*Body acceptance*. $n = 761$.

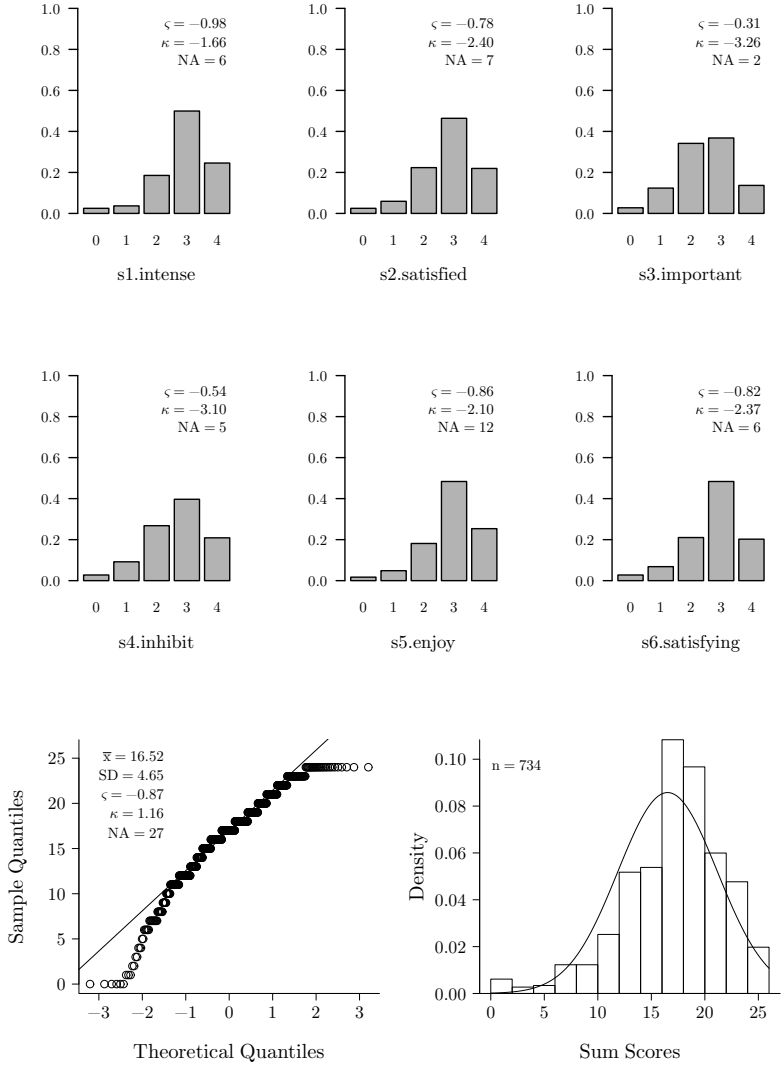



*Figure 7.3.* Item barplots and sum score Q-Q plot and density plot for DBIQ-*Sexual fulfillment.* $n = 761$.

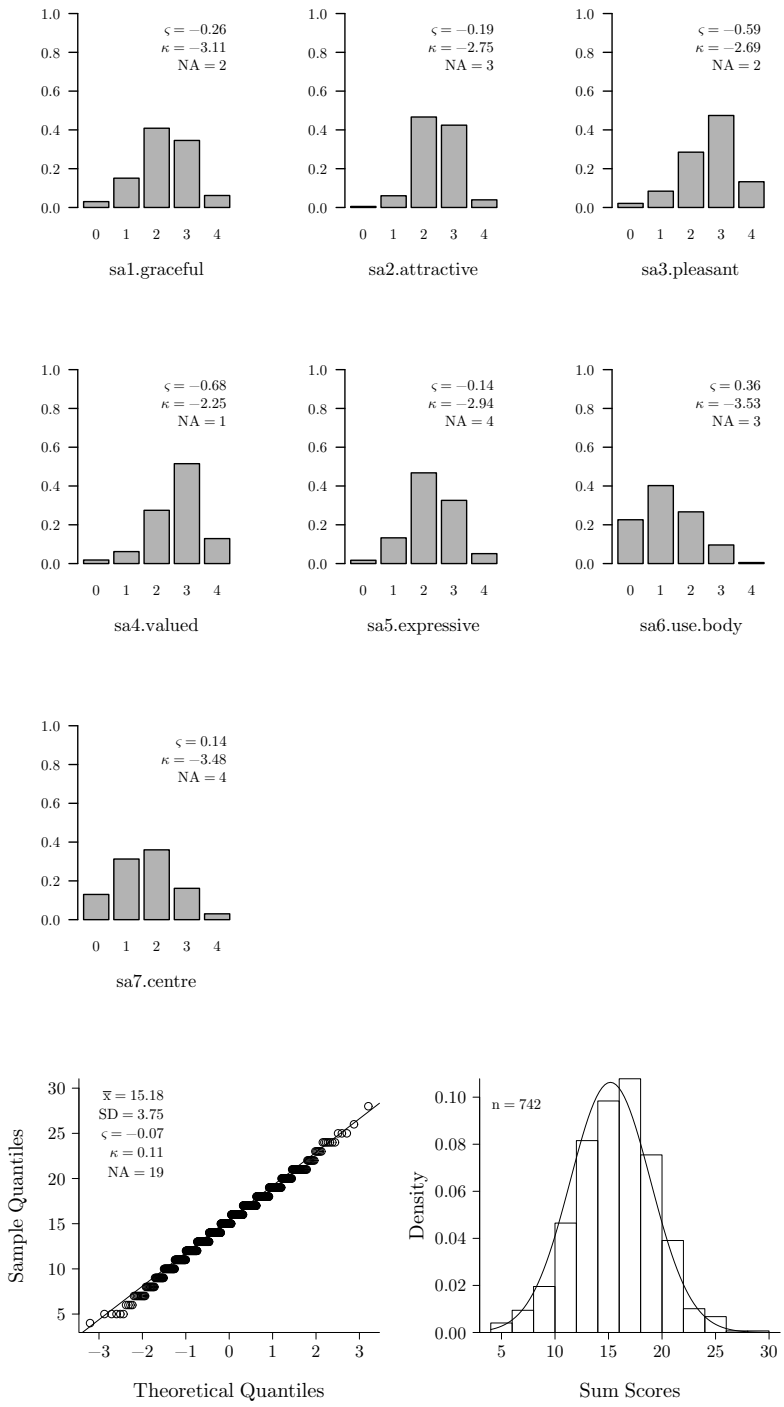



*Figure 7.4.* Item barplots and sum score Q-Q plot and density plot for DBIQ-*Self-aggrandizement.* $n = 761$.

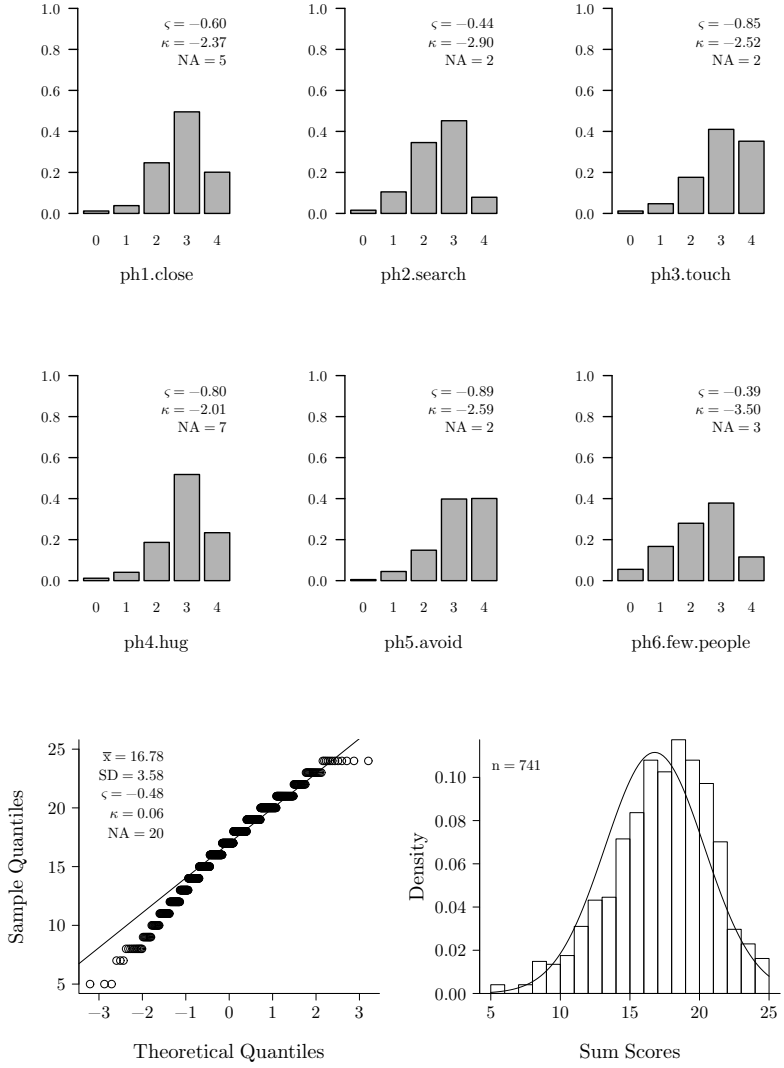



*Figure 7.5.* Item barplots and sum score Q-Q plot and density plot for DBIQ-*Physical contact.* $n = 761$.

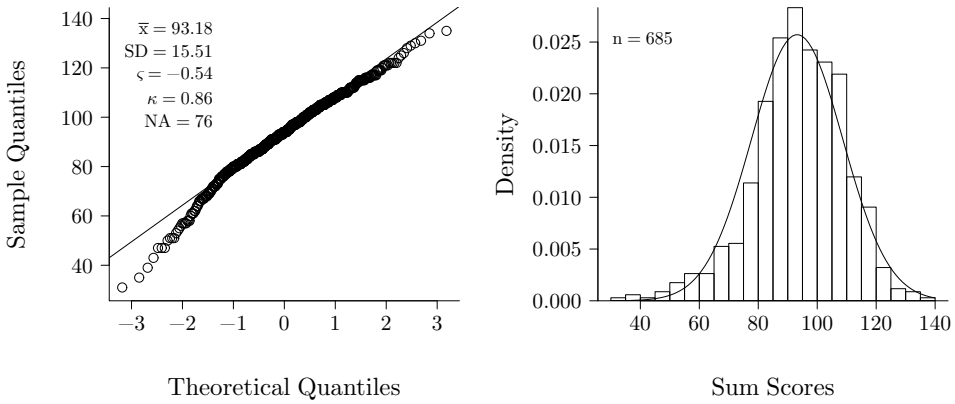



*Figure 7.6.* Sum score Q-Q plot and density plot for total DBIQ. $n = 761$.

The sum score distribution of the final subscale *Physical contact* is left-skewed ($\varsigma = -0.48$). All items constituting that scale are left-skewed with skewness between $-0.89$ and $-0.39$, and excess kurtosis ranging from $-0.50$ to 0.98.

In summary, most items deviate from normality, but not extremely. Summing the items leads to sum score distributions varying from normal to left-skewed.

### 7.3.2 Results

#### LV Score Estimation

For each subscale as well as the total scale, sum score distributions are given in Figures 7.1 to 7.6. Sum scores are the LV estimates in an IRT-mok analysis. In practice, however, sum scores are also often used as LV estimates when any of the parametric models are applied. We explicitly distinguish between FA-lin, FA-poly, IRT-grm, and IRT-mok LV estimates, with the former three being estimated from the model parameters and the latter being the sum scores. LV scores as estimated by FA-lin, FA-poly, and IRT-grm are given in Figures 7.7 to 7.11.

*Table 7.2.* Skewness of LV score estimates for DBIQ subscales; $n = 761$

| Subscale | LV skewness | | | |
|---|---|---|---|---|
| | IRT-mok | FA-lin | FA-poly | IRT-grm |
| *Vitality* | −0.34 | −0.41 | 0.14 | 0.14 |
| *Body acceptance* | −0.75 | −0.79 | −0.23 | −0.29 |
| *Sexual fulfillment* | −0.87 | −0.93 | −0.29 | −0.34 |
| *Self-aggrandizement* | −0.07 | −0.13 | 0.04 | 0.06 |
| *Physical contact* | −0.48 | −0.48 | 0.00 | −0.03 |

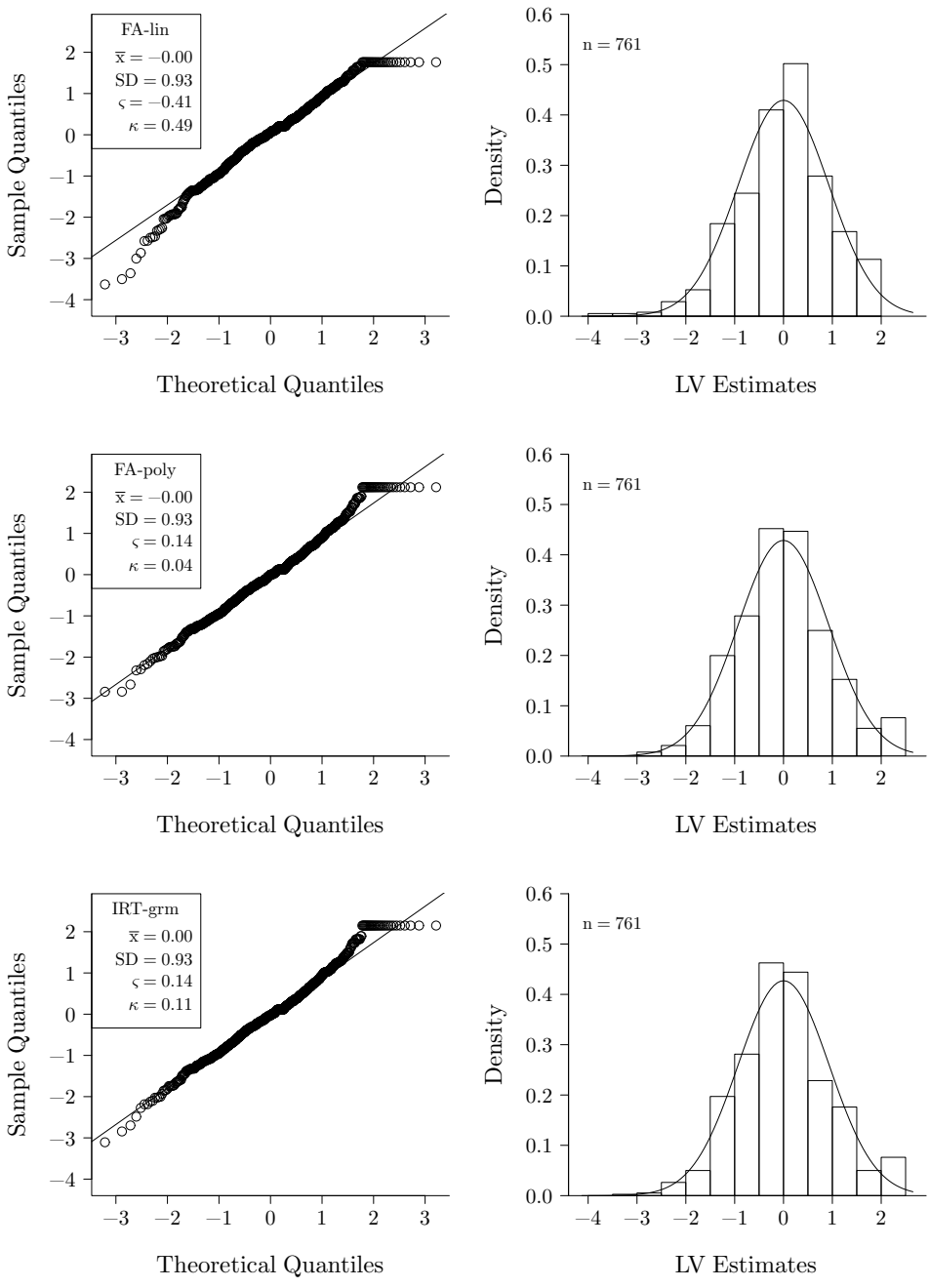



*Figure 7.7.* LV score distribution of DBIQ-*Vitality*. $n = 761$.

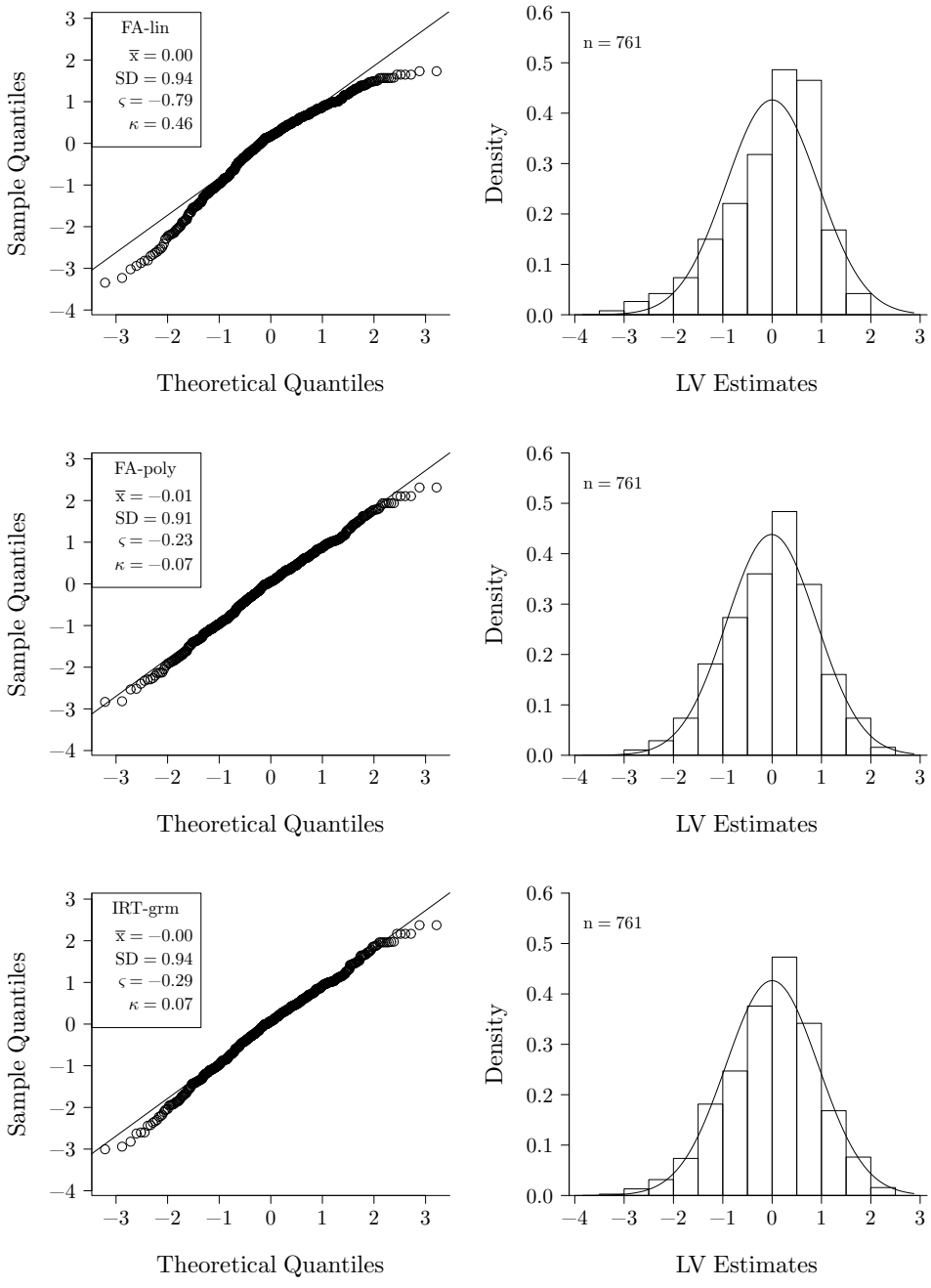



*Figure 7.8.* LV score distribution of DBIQ-*Body acceptance*. $n = 761$.

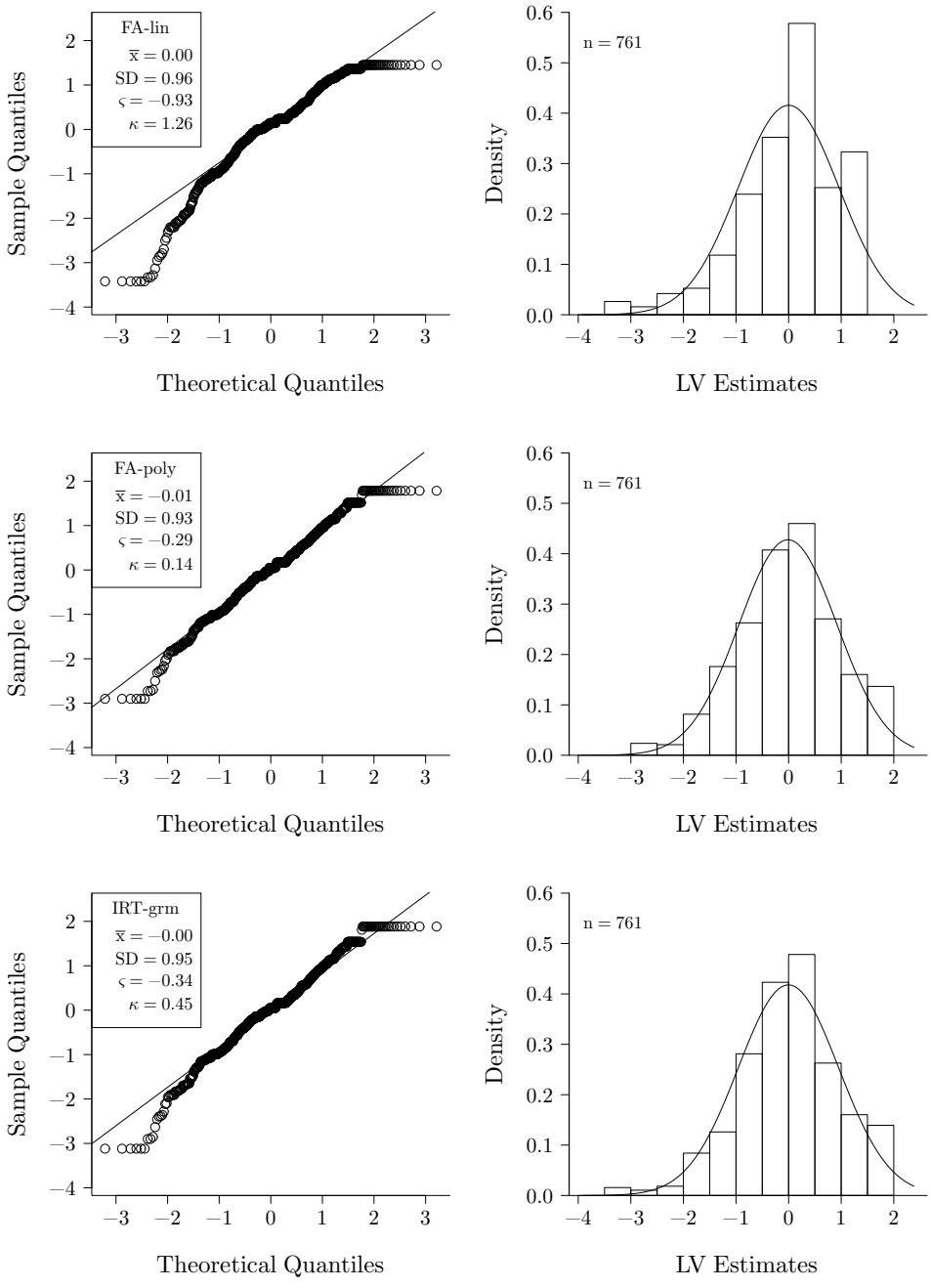



*Figure 7.9.* LV score distribution of DBIQ-*Sexual fulfillment*. $n = 761$.

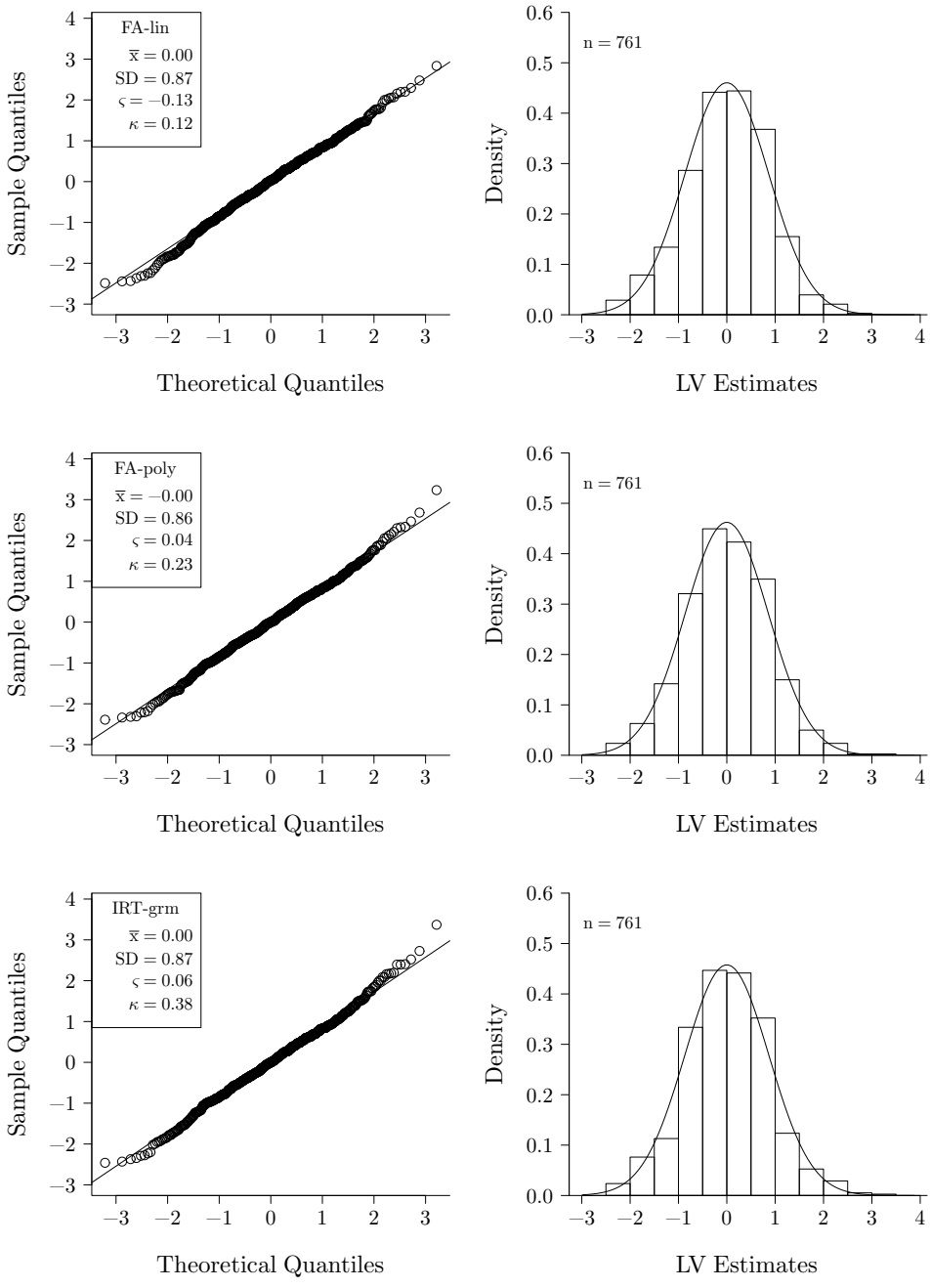



*Figure 7.10.* LV score distribution of DBIQ-*Self-aggrandizement.* $n = 761$.

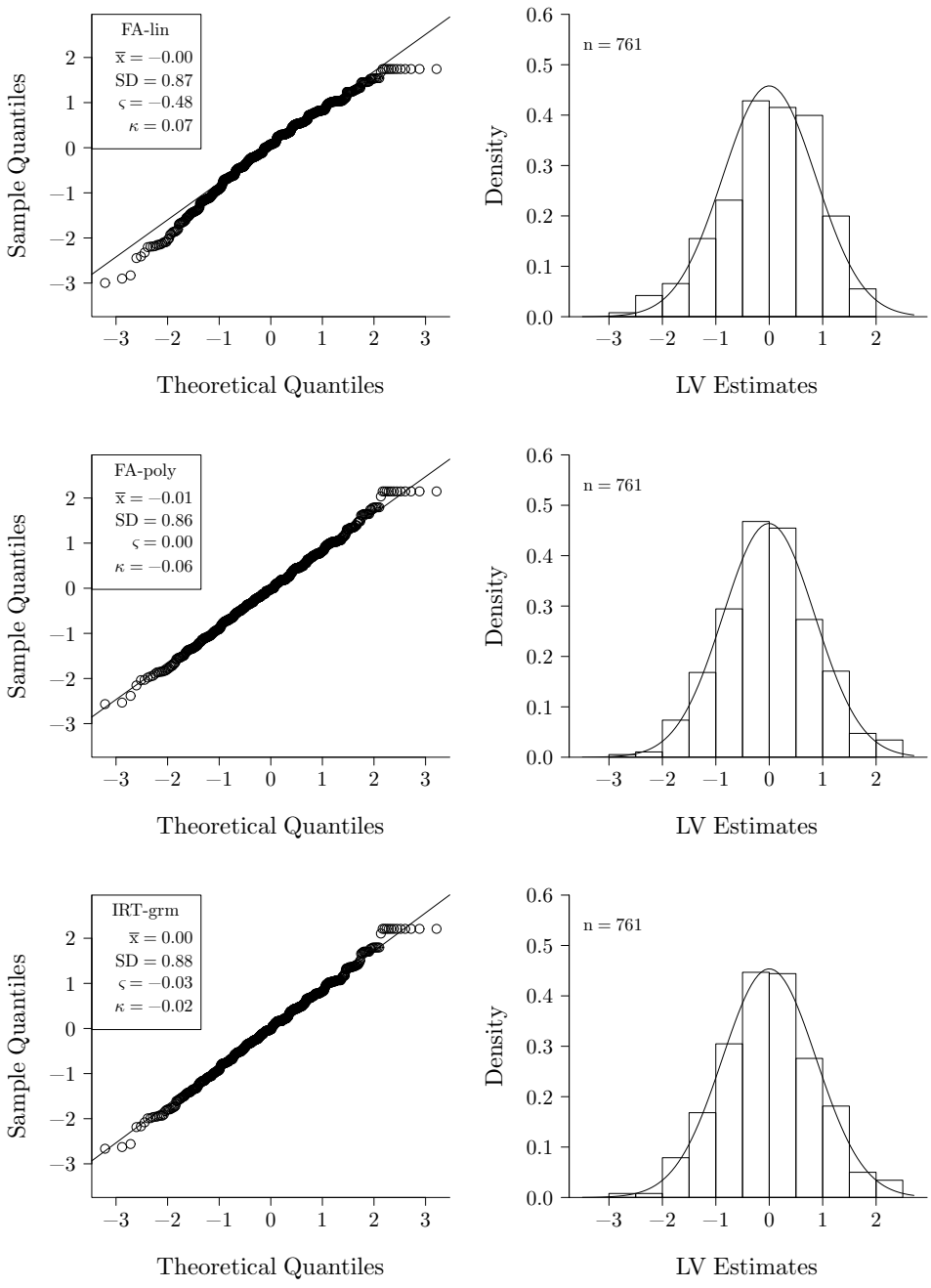



*Figure 7.11.* LV score distribution of DBIQ-*Physical contact*. $n = 761$.

In Table 7.2 the skewness of the LV estimates for the five subscales is presented. First notice the similarity of IRT-mok and FA-lin estimated skewness. FA-poly and IRT-grm results are also very similar, but notably different from the IRT-mok and FA-lin results, with regard to the sign and the size of the skewness. In our simulation study, we also found a distinction between IRT-mok and FA-lin versus FA-poly and IRT-grm for LV score estimation. Comparing the present findings to the results of our simulation study, we can make inferences about the true distributions of these LVs.

The first subscale *Vitality* has left-skewed item distributions, a left-skewed sum score distribution, and right-skewed FA-poly and IRT-grm LV distributions. This pattern neatly fits the case of a right-skewed LV combined with left-skewed items, i.e., Cell lnRS6 in our Monte Carlo design (see Table 6.1 on p. 145), except that for *Vitality* all items are left-skewed, whereas in our design half the items were left-skewed and half were normal.

The second and third subscales *Body acceptance* and *Sexual fulfillment* have mostly left-skewed items, a left-skewed sum score distribution, and left-skewed FA-poly/IRT-grm LV distributions, where the latter are less skewed than the sum score distribution. This pattern resembles that of a skewed LV distribution with items skewed in the same direction (cf. Cell rnRS6).

The fourth subscale *Self-aggrandizement* has a normal sum score distribution combined with a mix of approximately normal, mildly-skewed, and skewed item distributions. The FA-poly and IRT-grm LV distributions are normal. This pattern indicates a normal LV with a mix of item distributions, similar to Cell lrnN6.

The final subscale *Physical contact* consists of left-skewed items, has a left-skewed sum score distribution, and normal FA-poly/IRT-grm LV distributions. This pattern resembles Cell lnNS6, with left-skewed items loading on a normal LV.

These inferences about the LVs are interesting in themselves, as they provide information on the latent variables of interest. In addition, they are useful for the interpretation of additional model estimation results, as we will see in the next subsection.

### Parameter and Standard Error Estimation

In this section the parameter estimates and corresponding standard error estimates are discussed for each subscale, comparing the results of the three parametric models. The LV results are used for the interpretation of the estimates.

Loading parameter estimates and corresponding standard errors of the five subscales are presented in Table 7.3, which contains the results of the unidimensional as well as the multidimensional analysis. Threshold parameter estimates are given in Table 7.4 for the *Vitality* subscale. For the remaining subscales, the threshold parameter and standard error estimates are provided in Tables F.1 to F.4 of Appendix F. For now, we focus on the unidimensional results, while turning to the multidimensional case later.

As for the *Vitality* subscale, we notice the smaller FA-lin parameter estimates as compared to the FA-poly and IRT-grm estimates for each item. FA-poly and IRT-grm

*Table 7.3.* Loading parameter and standard error estimates for the DBIQ subscales *Vitality*, *Body acceptance*, *Sexual fulfillment*, *Self-aggrandizement*, and *Physical contact*. $n = 761$.

| | FA-lin | | FA-poly | | IRT-grm | |
|---|---|---|---|---|---|---|
| | $\hat{\lambda}_{uni}$ $(\hat{se})$ | $\hat{\lambda}_{multi}$ $(\hat{se})$ | $\hat{\lambda}_{uni}$ $(\hat{se})$ | $\hat{\lambda}_{multi}$ $(\hat{se})$ | $\hat{\lambda}_{uni}$ $(\hat{se})$ | $\hat{\lambda}_{multi}$ $(\hat{se})$ |
| v1.run.down | 0.549 (0.029) | 0.545 (0.029) | 0.635 (0.022) | 0.601 (0.027) | 0.628 (0.033) | 0.566 (0.032) |
| v2.motivation | 0.439 (0.032) | 0.446 (0.032) | 0.520 (0.026) | 0.564 (0.027) | 0.517 (0.035) | 0.783 (0.026) |
| v3.exhaust | 0.657 (0.024) | 0.652 (0.024) | 0.731 (0.019) | 0.696 (0.023) | 0.721 (0.028) | 0.791 (0.026) |
| v4.fit | 0.741 (0.020) | 0.748 (0.019) | 0.791 (0.017) | 0.809 (0.018) | 0.811 (0.021) | 0.775 (0.023) |
| v5.energy | 0.755 (0.019) | 0.746 (0.020) | 0.807 (0.015) | 0.788 (0.018) | 0.824 (0.020) | 0.503 (0.038) |
| v6.ph.cond | 0.736 (0.021) | 0.749 (0.020) | 0.799 (0.017) | 0.846 (0.017) | 0.803 (0.023) | 0.537 (0.040) |
| v7.ph.limits | 0.540 (0.029) | 0.537 (0.029) | 0.599 (0.025) | 0.596 (0.028) | 0.628 (0.034) | 0.798 (0.023) |
| v8.ph.strong | 0.652 (0.024) | 0.645 (0.024) | 0.729 (0.020) | 0.706 (0.024) | 0.730 (0.027) | 0.858 (0.016) |
| ba1.happy | 0.727 (0.020) | 0.729 (0.020) | 0.784 (0.018) | 0.763 (0.021) | 0.781 (0.024) | 0.695 (0.026) |
| ba2.like | 0.758 (0.019) | 0.782 (0.017) | 0.805 (0.016) | 0.847 (0.015) | 0.805 (0.022) | 0.830 (0.019) |
| ba3.clothing | 0.529 (0.029) | 0.527 (0.029) | 0.569 (0.027) | 0.562 (0.029) | 0.564 (0.032) | 0.947 (0.010) |
| ba4.uncomf | 0.723 (0.021) | 0.713 (0.021) | 0.787 (0.020) | 0.788 (0.022) | 0.785 (0.026) | 0.869 (0.017) |
| ba5.wish.diff | 0.762 (0.019) | 0.731 (0.020) | 0.829 (0.017) | 0.756 (0.021) | 0.817 (0.024) | 0.419 (0.041) |
| ba6.satisfied | 0.706 (0.021) | 0.713 (0.021) | 0.766 (0.019) | 0.763 (0.020) | 0.770 (0.024) | 0.676 (0.037) |
| ba7.change | 0.518 (0.029) | 0.479 (0.031) | 0.554 (0.027) | 0.428 (0.031) | 0.541 (0.034) | 0.709 (0.033) |
| ba8.show | 0.464 (0.031) | 0.511 (0.030) | 0.491 (0.028) | 0.667 (0.024) | 0.484 (0.040) | 0.580 (0.038) |
| s1.intense | 0.734 (0.019) | 0.743 (0.018) | 0.782 (0.016) | 0.816 (0.017) | 0.786 (0.023) | 0.520 (0.040) |
| s2.satisfied | 0.821 (0.014) | 0.817 (0.014) | 0.868 (0.010) | 0.854 (0.012) | 0.862 (0.016) | 0.493 (0.045) |
| s3.important | 0.649 (0.023) | 0.654 (0.022) | 0.688 (0.020) | 0.706 (0.021) | 0.690 (0.026) | 0.564 (0.036) |
| s4.inhibit | 0.769 (0.017) | 0.773 (0.016) | 0.812 (0.014) | 0.816 (0.015) | 0.826 (0.020) | 0.655 (0.039) |
| s5.enjoy | 0.896 (0.010) | 0.896 (0.010) | 0.936 (0.007) | 0.939 (0.008) | 0.945 (0.011) | 0.791 (0.035) |
| s6.satisfying | 0.833 (0.013) | 0.825 (0.014) | 0.872 (0.011) | 0.847 (0.013) | 0.877 (0.016) | 0.602 (0.044) |
| sa1.graceful | 0.374 (0.038) | 0.411 (0.035) | 0.404 (0.036) | 0.483 (0.035) | 0.378 (0.044) | 0.621 (0.034) |
| sa2.attractive | 0.545 (0.033) | 0.612 (0.029) | 0.595 (0.029) | 0.764 (0.027) | 0.603 (0.040) | 0.557 (0.049) |
| sa3.pleasant | 0.664 (0.030) | 0.660 (0.027) | 0.739 (0.026) | 0.693 (0.028) | 0.713 (0.036) | 0.543 (0.044) |
| sa4.valued | 0.509 (0.035) | 0.514 (0.032) | 0.568 (0.028) | 0.583 (0.032) | 0.560 (0.041) | 0.655 (0.029) |
| sa5.expressive | 0.476 (0.035) | 0.480 (0.033) | 0.514 (0.032) | 0.520 (0.035) | 0.516 (0.044) | 0.467 (0.038) |
| sa6.use.body | 0.547 (0.033) | 0.479 (0.034) | 0.582 (0.032) | 0.451 (0.035) | 0.575 (0.042) | 0.600 (0.032) |
| sa7.centre | 0.602 (0.031) | 0.553 (0.031) | 0.637 (0.027) | 0.530 (0.031) | 0.621 (0.035) | 0.332 (0.049) |
| ph1.close | 0.533 (0.037) | 0.597 (0.030) | 0.637 (0.026) | 0.631 (0.031) | 0.592 (0.054) | 0.555 (0.051) |
| ph2.search | 0.622 (0.034) | 0.737 (0.026) | 0.725 (0.022) | 0.827 (0.026) | 0.680 (0.050) | 0.583 (0.037) |
| ph3.touch | 0.630 (0.033) | 0.539 (0.034) | 0.689 (0.024) | 0.663 (0.030) | 0.685 (0.046) | 0.193 (0.048) |
| ph4.hug | 0.535 (0.034) | 0.542 (0.031) | 0.590 (0.026) | 0.558 (0.032) | 0.613 (0.040) | 0.376 (0.046) |
| ph5.avoid | 0.590 (0.033) | 0.509 (0.034) | 0.648 (0.027) | 0.641 (0.034) | 0.634 (0.050) | 0.551 (0.038) |
| ph6.few.people | 0.585 (0.032) | 0.513 (0.033) | 0.618 (0.026) | 0.556 (0.034) | 0.611 (0.044) | 0.662 (0.040) |

*Table 7.4.* Threshold parameter and standard error estimates for DBIQ-*Vitality*. $n = 761$.

| | FA-poly | | IRT-grm | |
|---|---|---|---|---|
| | $\hat{\tau}$ | $\hat{se}(\hat{\tau})$ | $\hat{\tau}$ | $\hat{se}(\hat{\tau})$ |
| $\tau_{1.1}$ | −2.480 | 0.159 | −2.592 | 0.197 |
| $\tau_{1.2}$ | −1.550 | 0.072 | −1.544 | 0.074 |
| $\tau_{1.3}$ | −0.721 | 0.050 | −0.708 | 0.047 |
| $\tau_{1.4}$ | 0.666 | 0.049 | 0.629 | 0.048 |
| $\tau_{2.1}$ | −2.355 | 0.140 | −2.460 | 0.176 |
| $\tau_{2.2}$ | −1.392 | 0.066 | −1.366 | 0.067 |
| $\tau_{2.3}$ | −0.575 | 0.048 | −0.572 | 0.045 |
| $\tau_{2.4}$ | 0.606 | 0.049 | 0.553 | 0.047 |
| $\tau_{3.1}$ | −2.790 | 0.229 | −2.902 | 0.272 |
| $\tau_{3.2}$ | −1.572 | 0.073 | −1.530 | 0.073 |
| $\tau_{3.3}$ | −0.641 | 0.049 | −0.623 | 0.046 |
| $\tau_{3.4}$ | 0.804 | 0.051 | 0.776 | 0.052 |
| $\tau_{4.1}$ | −2.150 | 0.114 | −2.147 | 0.122 |
| $\tau_{4.2}$ | −1.658 | 0.077 | −1.633 | 0.075 |
| $\tau_{4.3}$ | −0.635 | 0.049 | −0.627 | 0.047 |
| $\tau_{4.4}$ | 0.779 | 0.051 | 0.748 | 0.052 |
| $\tau_{5.1}$ | −2.480 | 0.159 | −2.455 | 0.158 |
| $\tau_{5.2}$ | −1.607 | 0.075 | −1.552 | 0.073 |
| $\tau_{5.3}$ | −0.456 | 0.047 | −0.456 | 0.045 |
| $\tau_{5.4}$ | 0.918 | 0.053 | 0.896 | 0.055 |
| $\tau_{6.1}$ | −2.414 | 0.148 | −2.412 | 0.157 |
| $\tau_{6.2}$ | −1.773 | 0.084 | −1.723 | 0.084 |
| $\tau_{6.3}$ | −0.631 | 0.049 | −0.622 | 0.046 |
| $\tau_{6.4}$ | 0.810 | 0.051 | 0.782 | 0.053 |
| $\tau_{7.1}$ | −2.183 | 0.118 | −2.205 | 0.135 |
| $\tau_{7.2}$ | −1.215 | 0.060 | −1.184 | 0.058 |
| $\tau_{7.3}$ | −0.313 | 0.046 | −0.334 | 0.043 |
| $\tau_{7.4}$ | 0.969 | 0.054 | 0.922 | 0.054 |
| $\tau_{8.1}$ | −2.222 | 0.122 | −2.219 | 0.131 |
| $\tau_{8.2}$ | −1.633 | 0.076 | −1.588 | 0.076 |
| $\tau_{8.3}$ | −0.722 | 0.050 | −0.716 | 0.047 |
| $\tau_{8.4}$ | 0.718 | 0.050 | 0.682 | 0.050 |

parameter estimates are rather similar. Because the true LV distribution is presumed to be right-skewed and the item distributions are left-skewed (cf. Cell lnRS6), we expect an underestimation of loading parameters by all models, which is most severe for FA-lin and least severe for IRT-grm. Therefore, we take IRT-grm loading parameter estimates to be the best approximations to the true values.

Standard error estimates are larger for IRT-grm than for FA-lin and FA-poly. In our Monte Carlo study, we found standard errors to be underestimated by FA-lin and FA-poly, most severely for skewed items loading on a skewed LV, whereas IRT-grm standard error estimators were unbiased, except for normal items loading on a skewed LV. In addition, IRT-grm parameter estimators were found to be slightly less precise as compared to FA-lin and FA-poly in our simulation study. The combination of slightly smaller standard errors and their underestimation by FA-lin and FA-poly versus the slightly larger standard errors accurately estimated by IRT-grm could explain the

observed difference in standard error estimates. These findings regarding standard errors hold for each subscale.

Threshold parameter estimates differ somewhat between FA-poly and IRT-grm and, based on the findings of our simulation study, are presumed to be more accurate for IRT-grm than for FA-poly.

From the LV results, we concluded that *Body acceptance* is likely a left-skewed LV. As the item distributions are also left-skewed, we know that FA-lin parameter estimates are rather accurate, although expected to be slightly smaller than the true values. FA-poly loading parameters are presumably overestimated, whereas IRT-grm estimates are most accurate. The resulting expected pattern of FA-lin estimates being smaller than IRT-grm estimates, which are themselves smaller than FA-poly estimates, holds for most items. However, FA-poly and IRT-grm estimates are more similar to each other and more dissimilar from FA-lin estimates than in our simulation study, which could be caused by the fact that the skewness of these items is less extreme than in our generated data. The smaller item skewness might increase the negative bias of FA-lin estimates and decrease FA-poly's positive bias.

The properties of the *Sexual fulfillment* subscale are similar to those of *Body acceptance*, albeit that the skewness of the former is more severe. For *Sexual fulfillment*, FA-poly loading estimates are also quite similar to IRT-grm loading estimates, but oftentimes marginally smaller than IRT-grm.

*Self-aggrandizement* presumably has a normal distribution. The item distributions vary in shapes. As expected, FA-poly and IRT-grm loading parameter estimates are quite similar to each other, whereas FA-lin estimates are smaller for each item, as a result of the negative bias of FA-lin loading parameter estimators.

*Physical contact* is also assumed to be normally distributed, but the items of this scale are all left-skewed. Loading parameter results are similar to the previous subscale: FA-poly and IRT-grm estimates are relatively similar to each other and always larger than FA-lin estimates.

Over all subscales and estimation models, the item loadings vary between 0.374 and 0.945. We conclude that the subscales of the DBIQ are medium to strong, with *Sexual fullfilment* being the strongest scale and/or, presumably, the narrowest concept, since its item loadings are homogeneously high.

### Model Fit

To assess the fit of a model to the sample data, a variety of indices can be used. We present a number of fit indices from the output of the MPLUS computer program, as well as some additional fit measures computed with our own R code. In Tables 7.5 to 7.9 the indices are presented for each subscale, where the upper panel contains MPLUS output, the lower panel additional indices. Note that the root mean squared error of approximation (RMSEA) values from the upper and lower panel are based on the $\chi^2$ and the $\chi^2_{YB}$ values from corresponding panels, respectively. For FA-lin and IRT-grm the Akaike's information criterion (AIC) and Bayes' information criterion (BIC) fit indices are also computed by MPLUS. Since they can only be used to compare

*Table 7.5.* Model fit results for DBIQ-*Vitality*. $n = 761$.

| Fit statistic | FA-lin | FA-poly | IRT-grm |
|---|---|---|---|
| $\chi^2_{mplus}$ | 191.452 | 343.401 | 7182.141 |
| df | 20 | 20 | 390483 |
| RMSEA | 0.106 | 0.146 | |
| CFI | 0.916 | 0.943 | |
| TLI | 0.882 | 0.920 | |
| $\chi^2_{YB}$ | 88.680 | | |
| df | 20 | | |
| RMSEA | 0.067 | | |
| SRMR | 0.053 | 0.055 | 0.063 |

*Table 7.6.* Model fit results for DBIQ-*Body acceptance*. $n = 761$.

| Fit statistic | FA-lin | FA-poly | IRT-grm |
|---|---|---|---|
| $\chi^2_{mplus}$ | 160.854 | 237.972 | 7449.106 |
| df | 20 | 20 | 390457 |
| RMSEA | 0.096 | 0.120 | |
| CFI | 0.936 | 0.963 | |
| TLI | 0.910 | 0.948 | |
| $\chi^2_{YB}$ | 100.345 | | |
| df | 20 | | |
| RMSEA | 0.073 | | |
| SRMR | 0.046 | 0.046 | 0.047 |

*Table 7.7.* Model fit results for DBIQ-*Sexual fulfillment*. $n = 761$.

| Fit statistic | FA-lin | FA-poly | IRT-grm |
|---|---|---|---|
| $\chi^2_{mplus}$ | 43.180 | 73.449 | 2357.974 |
| df | 9 | 9 | 15561 |
| RMSEA | 0.071 | 0.097 | |
| CFI | 0.988 | 0.994 | |
| TLI | 0.979 | 0.990 | |
| $\chi^2_{YB}$ | 24.993 | | |
| df | 9 | | |
| RMSEA | 0.048 | | |
| SRMR | 0.021 | 0.026 | 0.027 |

nested models, they are not presented nor discussed for the DBIQ or for the other applications.

The first fit statistic provided in the tables is the $\chi^2$ from the MPLUS output (denoted in the tables by $\chi^2_{mplus}$). For FA-lin and FA-poly it is very large compared to the number of degrees of freedom for each subscale, which points to a lack of model fit. It should be noted that the $\chi^2_{mplus}$ is very sensitive to nonnormality, its value being inflated, and it is a direct function of the sample size. The RMSEA is

*Table 7.8.* Model fit results for DBIQ-*Self-aggrandizement.* $n = 761$.

| Fit statistic | FA-lin | FA-poly | IRT-grm |
|---|---|---|---|
| $\chi^2_{mplus}$ | 125.938 | 172.806 | 5789.467 |
| df | 14 | 14 | 78006 |
| RMSEA | 0.103 | 0.122 | |
| CFI | 0.876 | 0.909 | |
| TLI | 0.814 | 0.864 | |
| $\chi^2_{YB}$ | 78.474 | | |
| df | 14 | | |
| RMSEA | 0.078 | | |
| SRMR | 0.058 | 0.059 | 0.058 |

*Table 7.9.* Model fit results for DBIQ-*Physical contact.* $n = 761$.

| Fit statistic | FA-lin | FA-poly | IRT-grm |
|---|---|---|---|
| $\chi^2_{mplus}$ | 178.876 | 279.367 | 3728.391 |
| df | 9 | 9 | 15556 |
| RMSEA | 0.157 | 0.199 | |
| CFI | 0.828 | 0.883 | |
| TLI | 0.713 | 0.804 | |
| $\chi^2_{YB}$ | 62.275 | | |
| df | 9 | | |
| RMSEA | 0.088 | | |
| SRMR | 0.075 | 0.079 | 0.076 |

less affected by nonnormality, and should thus provide more reliable information on the fit of the model than the $\chi^2$. For all subscales, the RMSEA favors FA-lin over FA-poly. The CFI and the TLI, on the other hand, consistently indicate a closer fit of the FA-poly model to the data.

For IRT-grm, we note the large $\chi^2$ value and the corresponding number of degrees of freedom, which is many times larger than the FA-lin and FA-poly $\chi^2$ for each subscale. This is due to the fact that for IRT-grm — taking into account the full item responses rather than merely the bivariate information, as for FA-lin and FA-poly — the item cross table is very large, with a number of cells equal to the number of categories raised to the power of the number of items $C^I$.

The SRMR values of the three estimation models are quite similar for each subscale, indicating that the sample covariance or correlation matrix is reproduced equally well by each of the models. Ranging from 0.021 to 0.079 they are all indicative of a good model fit.

Furthermore, we notice that the $\chi^2_{YB}$ cannot be computed for FA-poly and IRT-grm for any of the subscales, as a result of the small number of items compared to the five response categories. The $\chi^2_{YB}$ can only be used when the FA-poly/IRT-grm model is identified given the sample covariance matrix. This condition is fulfilled when the number of items is at least twice the number of response categories (Maydeu-Olivares et al., 2011).

In general, the results are indicative of a moderate fit of all three models to the data.

### Multidimensional Analysis

Since *body image* is presumed to be a multidimensional construct, one could very well justify a multidimensional analysis where all five subscales are modeled simultaneously, taking into account the dependencies between the subscales (cf. Scheffers et al., 2013).

As mentioned before, the estimated loading parameters and corresponding standard errors from the multidimensional FA-lin, FA-poly, and IRT-grm analyses are presented in Table 7.3 in addition to the unidimensional results.

The most noticeable differences between the unidimensional and multidimensional analyses occur for IRT-grm, where the uni- and multidimensional loading parameter estimates deviate most, with a maximum difference of 0.492 for Item *ph3.touch*. Standard error estimates, however, do not differ substantially.

The dissimilarities in parameter estimates are related to discrepancies in estimated correlations (or standardized covariances) $\phi$ of the LVs between the models, presented in Table 7.10. Large differences can be observed between FA-lin/FA-poly and IRT-grm. In most cases, IRT-grm estimates are larger, with a maximum difference for *Self-aggrandizement* and *Vitality* at 0.331/0.318 for FA-lin/FA-poly and 0.819 for IRT-grm. Standard error estimates do not seem to differ much between the models. It is further noted that the correlations between the unidimensional LV scores (not presented in

*Table 7.10.* Multivariate LV correlation parameter and standard error estimates for DBIQ. $n = 761$.

| | FA-lin | | FA-poly | | IRT-grm | |
|---|---|---|---|---|---|---|
| | $\hat{\phi}$ | $\hat{se}(\hat{\phi})$ | $\hat{\phi}$ | $\hat{se}(\hat{\phi})$ | $\hat{\phi}$ | $\hat{se}(\hat{\phi})$ |
| $\phi_{ba.v}$ | 0.626 | 0.028 | 0.651 | 0.023 | 0.625 | 0.033 |
| $\phi_{s.v}$ | 0.439 | 0.034 | 0.454 | 0.031 | 0.524 | 0.035 |
| $\phi_{s.ba}$ | 0.579 | 0.029 | 0.599 | 0.025 | 0.717 | 0.028 |
| $\phi_{sa.v}$ | 0.331 | 0.042 | 0.318 | 0.037 | 0.819 | 0.020 |
| $\phi_{sa.ba}$ | 0.553 | 0.036 | 0.633 | 0.027 | 0.815 | 0.020 |
| $\phi_{sa.s}$ | 0.555 | 0.033 | 0.595 | 0.028 | 0.814 | 0.022 |
| $\phi_{ph.v}$ | 0.181 | 0.044 | 0.216 | 0.038 | 0.625 | 0.033 |
| $\phi_{ph.ba}$ | 0.374 | 0.040 | 0.391 | 0.034 | 0.722 | 0.027 |
| $\phi_{ph.s}$ | 0.534 | 0.033 | 0.539 | 0.028 | 0.787 | 0.023 |
| $\phi_{ph.sa}$ | 0.646 | 0.035 | 0.614 | 0.029 | 0.826 | 0.021 |

the text) were very similar for all estimation models (close to the FA-lin and FA-poly multivariate LV correlations) and smaller than the multivariate correlations.

The reasons for the discrepancies in parameter estimates are not clear. They could, for example, lie in the fact that — in order to keep computation times manageable — the number of integration points for the multidimensional IRT-grm analysis was chosen to be smaller than for the unidimensional cases (9 versus 15). An increase of the number of integration points to 10, however, did not change anything in the multidimensional results. Further increase was not attempted, because of the lack of improvement resulting from the initial increase combined with the exponential increase of computation time.

### Analysis by IRT-mok

In addition to the parametric analyses, we applied a nonparametric IRT-mok analysis to each of the five subscales. For each subscale, the resulting Loevinger's item and scale $H$ coefficients and corresponding standard errors are presented in Table 7.11. Missing values were omitted listwise preceding the analysis.

The assumption of monotonicity was sufficiently met for each item and all subscales. All $H_{scale}$ values exceed the commonly applied lower-bound criterion of 0.30. In line with the parametric results, *Sexual fulfillment* is the strongest scale. The first item of the *Self-aggrandizement* subscale, *sa1.graceful*, is the weakest ($H_i = 0.231$), which also complies to our previously discussed findings.

The item response plots used in IRT-mok to investigate the monotonicity assumption also reveal the discriminative power of an item. As Item *sa1.graceful* seems to be a (relatively) malfunctioning item, we provide its item response plot in Figure 7.12. The horizontal axis represents the LV score, by means of the item rest score, which is the categorized scale sum score excluding Item *sa1.graceful*. The vertical axis represents the proportion of endorsement. The dashed lines are item-step response functions (ISRFs), i.e., $P(X_i \geq c)$ for each rest score group, with one line per cate-

*Table 7.11.* IRT-mok results for the DBIQ subscales *Vitality* ($n = 750$), *Body acceptance* ($n = 741$), *Sexual fulfillment* ($n = 734$), *Self-aggrandizement* ($n = 742$), and *Physical contact* ($n = 741$).

| Item | $\hat{H}$ | $\hat{se}(\hat{H})$ |
|---|---|---|
| v1.run.down | 0.402 | 0.026 |
| v2.motivation | 0.336 | 0.029 |
| v3.exhaust | 0.444 | 0.022 |
| v4.fit | 0.471 | 0.023 |
| v5.energy | 0.498 | 0.022 |
| v6.ph.cond | 0.475 | 0.022 |
| v7.ph.limits | 0.409 | 0.030 |
| v8.ph.strong | 0.433 | 0.025 |
| $\hat{H}_{scale}$ *Vitality* | 0.432 | 0.019 |
| ba1.happy | 0.528 | 0.021 |
| ba2.like | 0.552 | 0.022 |
| ba3.clothing | 0.389 | 0.027 |
| ba4.uncomf | 0.518 | 0.024 |
| ba5.wish.diff | 0.557 | 0.021 |
| ba6.satisfied | 0.510 | 0.023 |
| ba7.change | 0.360 | 0.028 |
| ba8.show | 0.340 | 0.030 |
| $\hat{H}_{scale}$ *Body acceptance* | 0.463 | 0.019 |
| s1.intense | 0.640 | 0.029 |
| s2.satisfied | 0.681 | 0.022 |
| s3.important | 0.592 | 0.028 |
| s4.inhibit | 0.657 | 0.025 |
| s5.enjoy | 0.750 | 0.018 |
| s6.satisfying | 0.688 | 0.023 |
| $\hat{H}_{scale}$ *Sexual fulfillment* | 0.668 | 0.021 |
| sa1.graceful | 0.231 | 0.026 |
| sa2.attractive | 0.358 | 0.024 |
| sa3.pleasant | 0.383 | 0.022 |
| sa4.valued | 0.299 | 0.025 |
| sa5.expressive | 0.297 | 0.026 |
| sa6.use.body | 0.332 | 0.026 |
| sa7.centre | 0.354 | 0.023 |
| $\hat{H}_{scale}$ *Self-aggrandizement* | 0.322 | 0.018 |
| ph1.close | 0.353 | 0.027 |
| ph2.search | 0.427 | 0.027 |
| ph3.touch | 0.407 | 0.026 |
| ph4.hug | 0.357 | 0.028 |
| ph5.avoid | 0.386 | 0.026 |
| ph6.few.people | 0.373 | 0.026 |
| $\hat{H}_{scale}$ *Physical contact* | 0.384 | 0.021 |

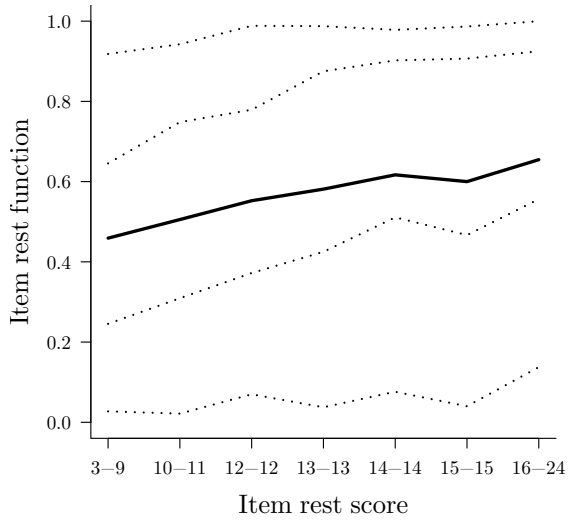



*Figure 7.12.* Item response plot for *sa1.graceful* based on IRT-mok analysis. The item rest score is based on the subscale items. The dashed lines represent the item step response functions $P(X_i \geq c)$ for $c = 0, \ldots, 3$. The solid line depicts the mean response function. $n = 742$.

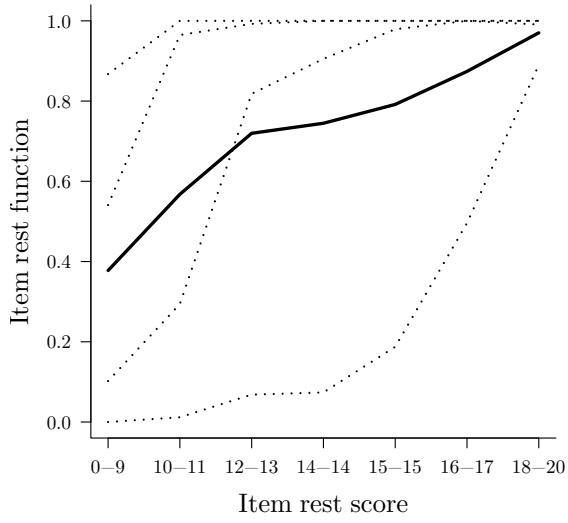



*Figure 7.13.* Item response plot for *s5.enjoy* based on IRT-mok analysis. The item rest score is based on the subscale items. The dashed lines represent the item step response functions $P(X_i \geq c)$ for $c = 0, \ldots, 3$. The solid line depicts the mean response function. $n = 734$.

gory $c = \{0, 1, 2, 3\}$, excluding the maximum category, as $P(X_i \geq 4) = 0$ holds by definition, regardless of the item rest score. The average of the four dotted lines is depicted with the solid line. When the assumption of monotonicity holds, the ISRFs are all monotone increasing. From Figure 7.12 we conclude for Item *sa1.graceful* that the monotonicity assumption approximately holds. The item's lack of discrimination, reflected both in the relatively small $H_i$ and $\lambda_i$ (for the parametric models), becomes apparent from the flat response lines: The summed score on the other items in the scale (i.e., the item rest score) is not strongly associated with the item score.

For comparative purposes, the item response plot of *s5.enjoy*, one of the strongest items in the analysis with $H_i = 0.750$, is presented in Figure 7.13. Here we see steep ISRFs, demonstrating the strong relation between this item score and the sum of the remaining items.

### 7.3.3 Discussion

We can draw a number of conclusions from the four employed analyses. First, the five subscales are moderate to strong, i.e., the items composing each subscale are associated sufficiently to form a scale of a single concept.

Second, from the parametric analyses, we can speculate about the characteristics of the LV distribution of each dimension of *body image*. Some of the dimensions are probably left-skewed, some right-skewed, and some are presumed to follow a normal distribution. Because the five subscales are not homogeneous, it is useful to regard the corresponding aspects of *body image* separately. Taken together, these subscales counterbalance each other, resulting in a more or less normal distribution, thus smoothing the differences of the components.

Third, from the brief examination of the multidimensional results, we conclude that, for FA-lin and FA-poly, results of five unidimensional analyses versus one five-dimensional analysis are quite similar. For IRT-grm, however, we found a number of large differences in loading estimates between the unidimensional and multidimensional analyses. The reasons for these deviations are not clear, and additional research is needed there. It is remarkable that the IRT-grm LV correlations are indicative of high consistency of the subscales, whereas the LV correlations as estimated by FA-lin and FA-poly are smaller and thus underline the multidimensionality of the scale more.

Finally, from the nonparametric IRT-mok analyses, we come to conclusions similar to those based upon our parametric analyses, with scaling coefficients $H_i$ quite comparable to the loading parameters.

So which estimation method is to be preferred here? Based on our simulation study, the FA-lin loading parameter estimates are likely to be too small. The FA-poly or IRT-grm parameter estimates are to be trusted most, with a slight preference for IRT-grm, since parameter bias was found to be smallest for this model in case of skewed LVs. Standard errors are also expected to be estimated more accurately by IRT-grm. Notably, the overall conclusions with regard to the quality of the scales based on either FA-poly or IRT-grm are generally the same. In addition, IRT-mok could equally well be used for analyzing the scale, and it produced results leading to conclusions

similar to those resulting from the FA-poly and IRT-grm applications. However, IRT-mok only provides an ordering of respondents on the LV. FA-poly and IRT-grm LV scores are better used for inferences on the approximate shape of the true LV distribution.

## 7.4 Revised Anticipated Sexual Jealousy Scale

The second scale under investigation is the Revised Anticipated Sexual Jealousy Scale (RASJS; Buunk, 1982, 1997), designed to measure the extent to which a respondent anticipates *"a negative affective response to various intimate and sexual behaviors of the partner"* (Buunk, 1997, p. 999). Originally a five-item scale measuring *reactive jealousy*, the RASJS was later extended to 15 items to cover two additional subtypes of jealousy: *anxious* and *possessive* jealousy.

The concept of sexual jealousy has been related to various aspects of relationship quality. Although in most research an increase in jealousy was associated with a decrease in relationship quality (e.g., Andersen, Eloy, Guerrero, & Spitzberg, 1995; Puente & Cohen, 2003; Shackelford & Buss, 2000), positive effects of jealousy on this outcome measure have also been identified (e.g., Barelds & Dijkstra, 2006). By taking a multidimensional approach of the concept of sexual jealousy, these seemingly contradictory results can be partly explained, with *reactive jealousy* being positively related to relationship quality, whereas *anxious* — and to a lesser extent *possessive* — jealousy are negatively associated with relationship quality (Barelds & Barelds-Dijkstra, 2007).

Here, we analyze the community sample ($n = 1366$, 683 women, 683 men) from Barelds and Dijkstra (2003), employing the FA and IRT scaling models under investigation. Respondents were randomly selected from phonebooks of cities throughout the Netherlands and sent an invitation to participate by regular mail. To participate they had to be married or living together as a couple (see Barelds and Dijkstra for a more detailed description of the sample).

Although the concept of sexual jealousy entails three related, but distinct dimensions, we analyze the composite scale of 15 items unidimensionally to gain insight in the scaling models' results when a multidimensional concept is misspecified as being unidimensional — or rather: oversimplified by regarding it at one level higher. We also employ three separate unidimensional analyses, taking into account the multidimensionality of the concept of sexual jealousy as perceived in the literature.

The description of the data in the next subsection is followed by the presentation of the results of the model applications and a discussion of our findings regarding the RASJS.

### 7.4.1 Descriptive Statistics

For 21 respondents, missing values were observed on one or more items. We removed these cases, resulting in a sample of $n = 1345$ for our analyses. The items of the RASJS and their means and standard deviations are given in Table 7.12.

*Table 7.12.* Revised Anticipated Sexual Jealousy Scale items ordered by subscales *Reactive*, *Anxious*, and *Possessive jealousy*.

| Abbreviation | Item Description | Mean[a] | SD |
|---|---|---|---|
| | How would you feel if your partner would ... with someone else? | | |
| r1.flirt | ... flirt ... | 2.24 | 1.38 |
| r2.discuss | ... discuss personal things ... | 1.94 | 1.46 |
| r3.sex.contact | ... have sexual contact ... | 3.82 | 0.58 |
| r4.dance | ... dance intimately ... | 2.57 | 1.37 |
| r5.kiss | ... kiss ... | 1.93 | 1.52 |
| a1.attractive | I am concerned about my partner finding someone else more attractive than me. | 0.71 | 0.89 |
| a2.sex.rel | I worry about the idea that my partner could have a sexual relationship with someone else. | 0.23 | 0.61 |
| a3.sex.int | I am afraid that my partner is sexually interested in someone else. | 0.27 | 0.60 |
| a4.general | I am concerned about all the things that could happen if my partner meets members of the opposite sex. | 0.31 | 0.65 |
| a5.leave | I worry that my partner might leave me for someone else. | 0.35 | 0.62 |
| p1.meet | I don't want my partner to meet too many people of the opposite sex. | 0.24 | 0.59 |
| p2.friend | It is not acceptable to me if my partner sees people of the opposite sex on a friendly basis. | 0.29 | 0.64 |
| p3.look | I demand from my partner that he/she does not look at other women/men. | 0.23 | 0.61 |
| p4.possessive | I am quite possessive with respect to my partner. | 0.74 | 1.10 |
| p5.own.way | I find it hard to let my partner go his/her own way. | 0.52 | 0.91 |

*Note.* The item descriptions were taken from Buunk (1997). The first subscale's five response categories varied from *"Not at all upset"* to *"Extremely upset"*, for the second subscale they varied from *"Never"* to *"Very often"*, and for the third subscale they varied from *"Not applicable"* to *"Very much applicable"*, all coded from 0 and 4.

[a] Item means and standard deviations were computed from the sample of size $n = 1345$.

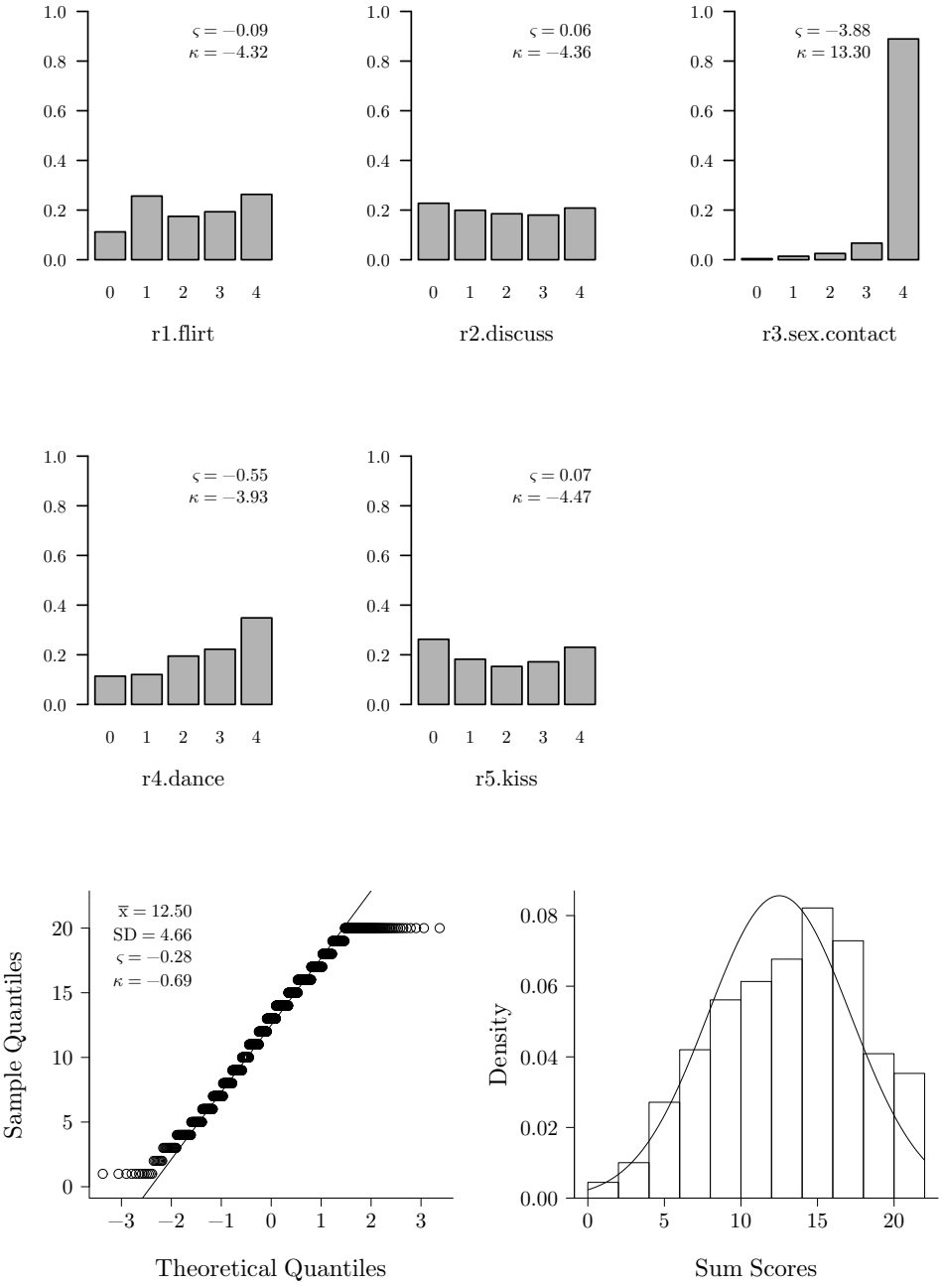



*Figure 7.14.* Item barplots and sum score Q-Q plot and density plot for RASJS-*Reactive jealousy.* $n = 1345$.

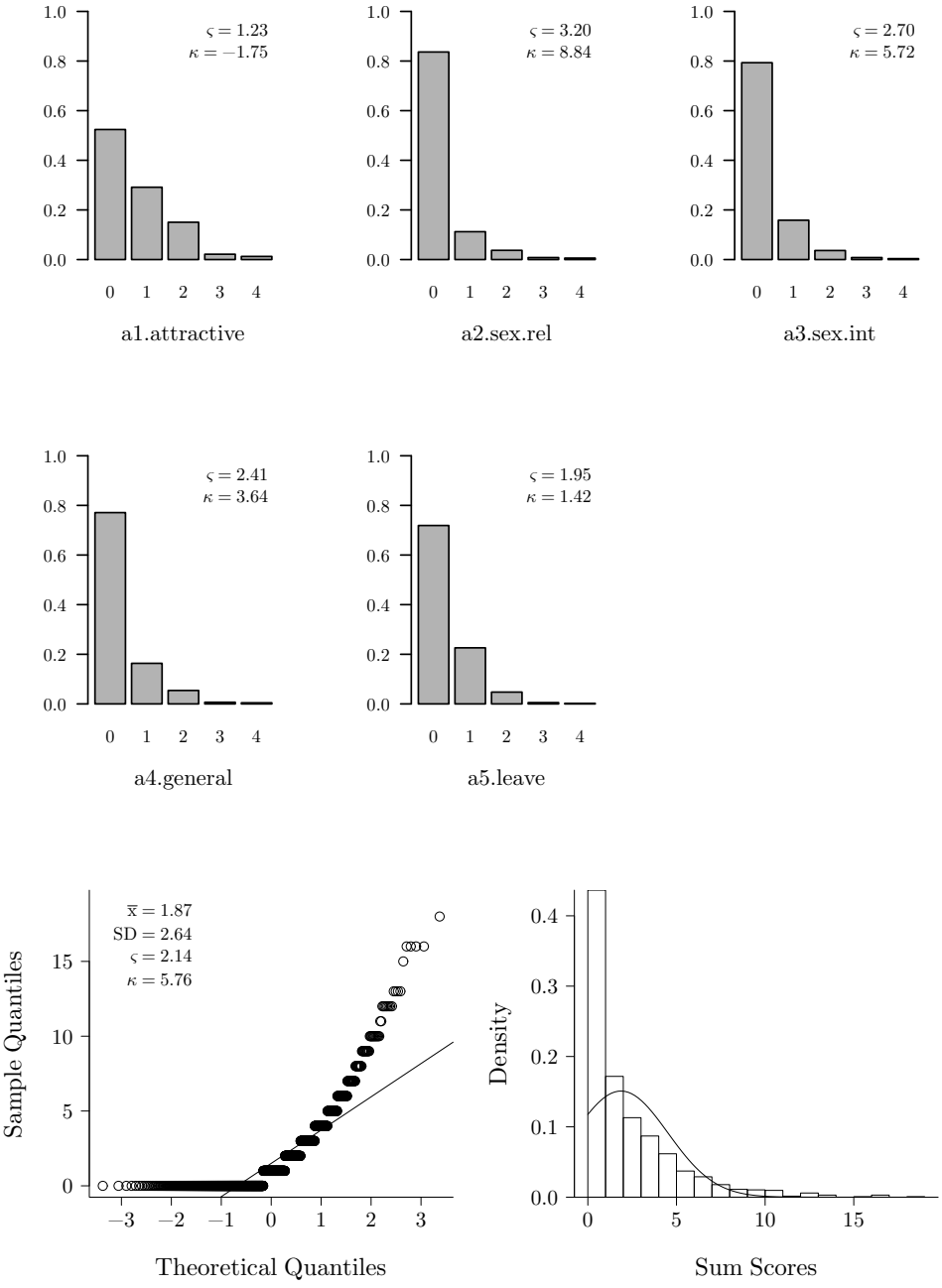



*Figure 7.15.* Item barplots and sum score Q-Q plot and density plot for RASJS-*Anxious jealousy.* $n = 1345$.

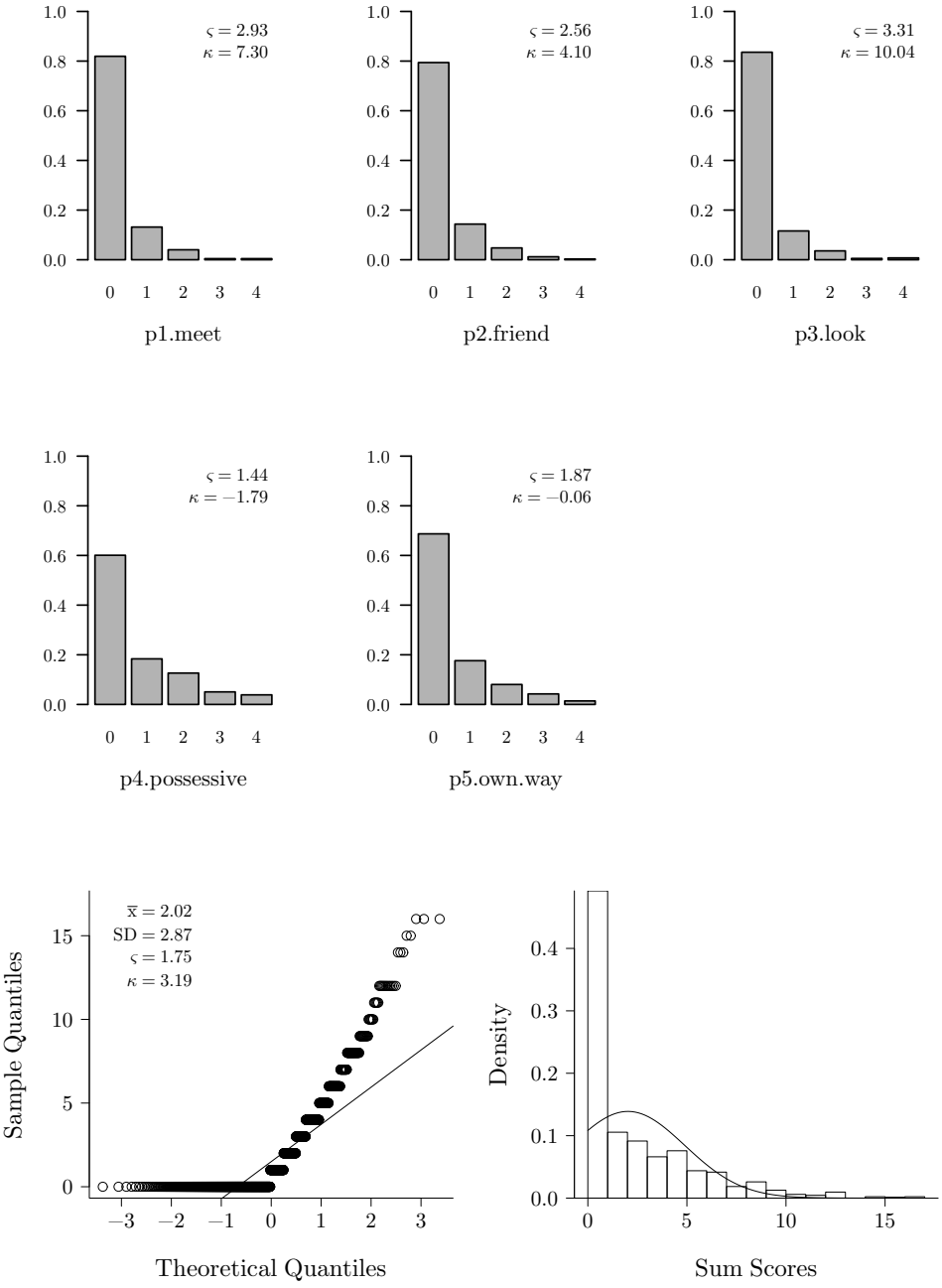



*Figure 7.16.* Item barplots and sum score Q-Q plot and density plot for RASJS-*Possessive jealousy.* $n = 1345$.

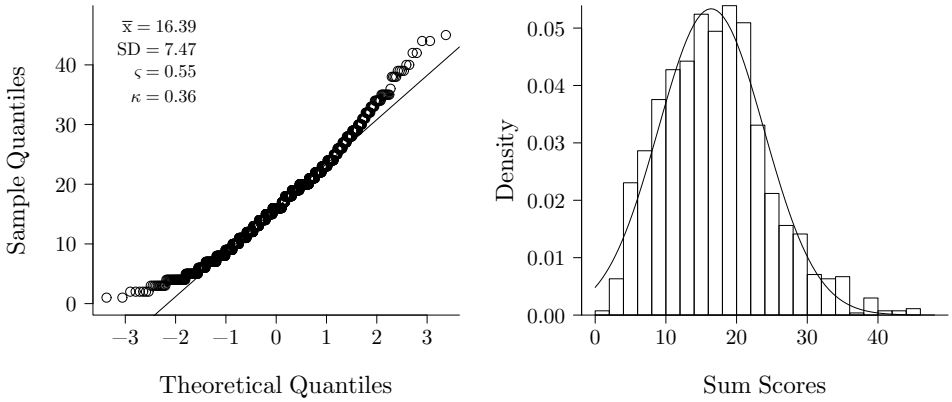



*Figure 7.17.* Sum score Q-Q plot and density plot for the total RASJS. $n = 1345$.

Graphs of the item and subscale distributions, and of the total score distribution can be found in Figures 7.14 to 7.17. In the inset of the item plots, values of skewness and excess kurtosis are provided. The sum score plots contain the empirical mean and standard deviation as well.

The most extreme item is *r3.sex.contact* of *Reactive jealousy*, which is highly left-skewed ($\varsigma = -3.88$, $\kappa = 16.24$). The distributional shapes of other items of that subscale are approximately uniform, left-skewed, and U-shaped.

The items of the *Anxious jealousy* and *Possessive jealousy* subscales are all right-skewed, with skewness ranging from 1.23 to 3.20, and 1.44 to 3.31, respectively. Excess kurtosis ranges from 1.24 to 11.79, and 1.20 to 12.99, respectively. The sum score distributions of both subscales are highly right-skewed ($\varsigma = 2.14$, $\kappa = 5.74$ for *Anxious jealousy*; and $\varsigma = 1.75$, $\kappa = 3.18$ for *Possessive jealousy*). Notice that on either scale many respondents score zero.

Overall, the item distributions of the RASJS are nowhere near normal. The extreme positive skewness encountered in the *Anxious jealousy* and *Possessive jealousy* subscales — many respondents score zero — is not uncommon for scales measuring a clinical LV employed in a community, or nonclinical, sample.

From Figure 7.17 it can be deduced that, when the three subscales are summed to the higher level of *Sexual jealousy*, the subscale distributions counterbalance each other to result in a moderately right-skewed sum score distribution ($\varsigma = 0.55$, $\kappa = 0.36$).

### 7.4.2 Results

#### LV Score Estimation

The scale distributions were given in Figures 7.14 to 7.17. In Figures 7.18 to 7.21 the LV scores as estimated by FA-lin, FA-poly, and IRT-grm are presented. Skewness information of all LV distributions is collected in Table 7.13, which we use to postulate the most plausible true shapes of the LV distributions.

Starting with the subscales *Anxious jealousy* and *Possessive jealousy*, consisting of right-skewed items only, all estimated LV score distributions are right-skewed, although IRT-mok (sum scores) and FA-lin are indicative of more severe skewness than FA-poly and IRT-grm. This pattern matches that of a right-skewed LV with right-skewed items, similar to those simulated in Cell rnRS6 (see Table 6.1 on p. 145) albeit that not all but half of the items were right-skewed there (the other half were normally distributed).

For subscale *Reactive jealousy* all LV distributions are estimated to be moderately left-skewed. Item distributions are approximately uniform, approximately bimodal, skewed, and extremely skewed. Such a pattern does not match any of the cells in our Monte Carlo design, but the closest resemblence is with Cell lrnRS6, i.e., items of mixed distributional shapes loading on a skewed LV. *Reactive jealousy* is thus most likely a left-skewed LV.

For the total RASJS, taking all items as a unidimensional scale, IRT-mok and FA-lin LV estimates are right-skewed. The magnitude of the skewness is quite dissimilar, presumably because the item loadings — which are discussed in more detail in the next subsection — are quite diverse in this scale and FA-lin takes these into account in the LV scores, whereas IRT-mok does not. FA-poly and IRT-grm produce normal LV distributions. Given that most item distributions are highly right-skewed, with some additional uniform and left-skewed items, the scale does not match any of the cells of our simulation study design. The greatest resemblance is arguably to Cell lrnNS6, although the loadings are mixed and not all strong. We therefore conclude that the total RASJS may best be represented by a normal LV.

*Table 7.13.* Skewness of LV score estimates for RASJS total and subscales. $n = 1345$.

| Subscale | LV Skewness | | | |
|---|---|---|---|---|
| | IRT-mok | FA-lin | FA-poly | IRT-grm |
| *Reactive jealousy* | −0.28 | −0.41 | −0.15 | −0.12 |
| *Anxious jealousy* | 2.14 | 2.52 | 0.83 | 0.78 |
| *Possessive jealousy* | 1.75 | 2.16 | 0.83 | 0.76 |
| RASJS total | 0.55 | 1.75 | 0.01 | 0.03 |

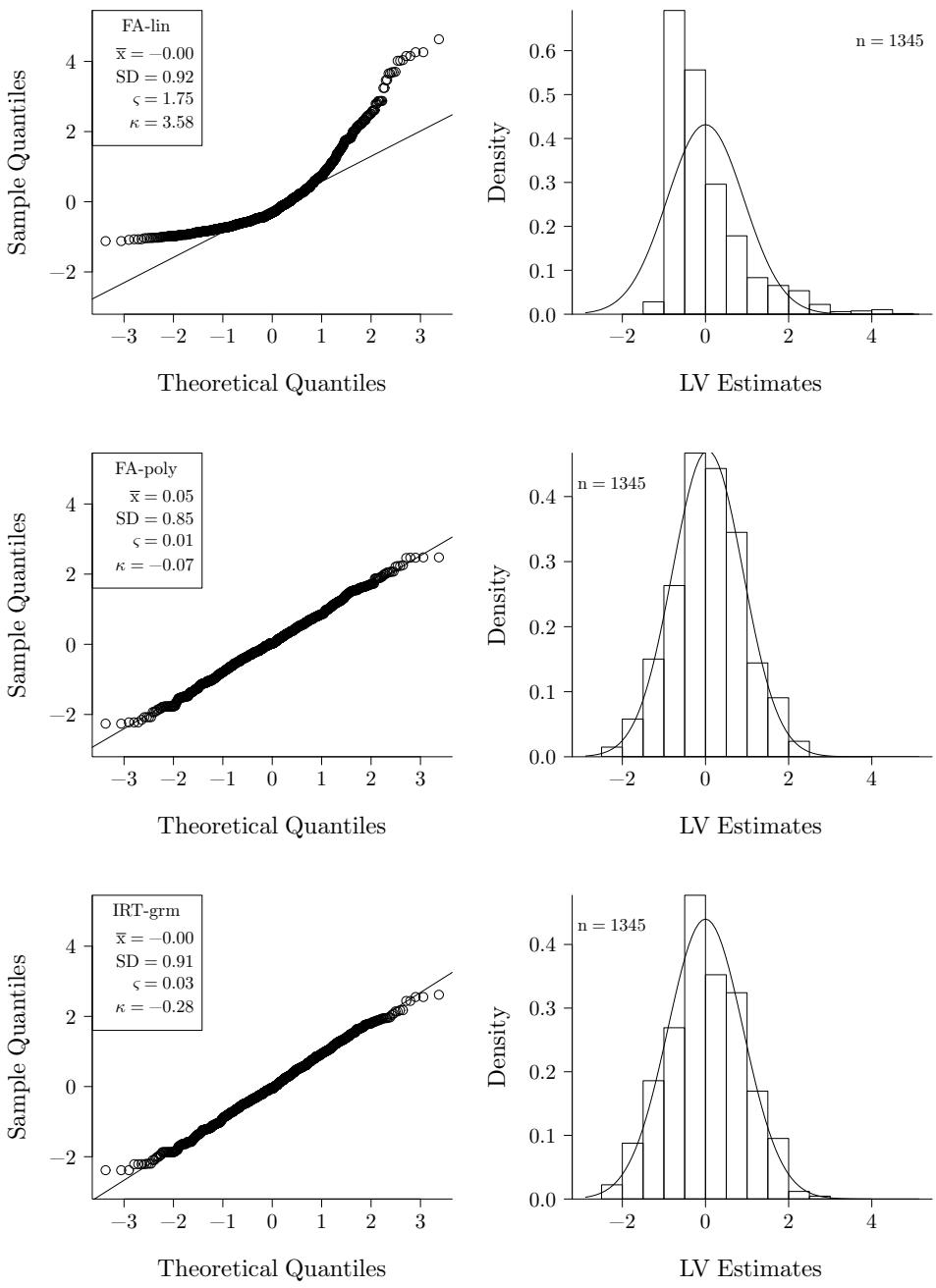



*Figure 7.18.* LV score distribution for total RASJS. $n = 1345$.

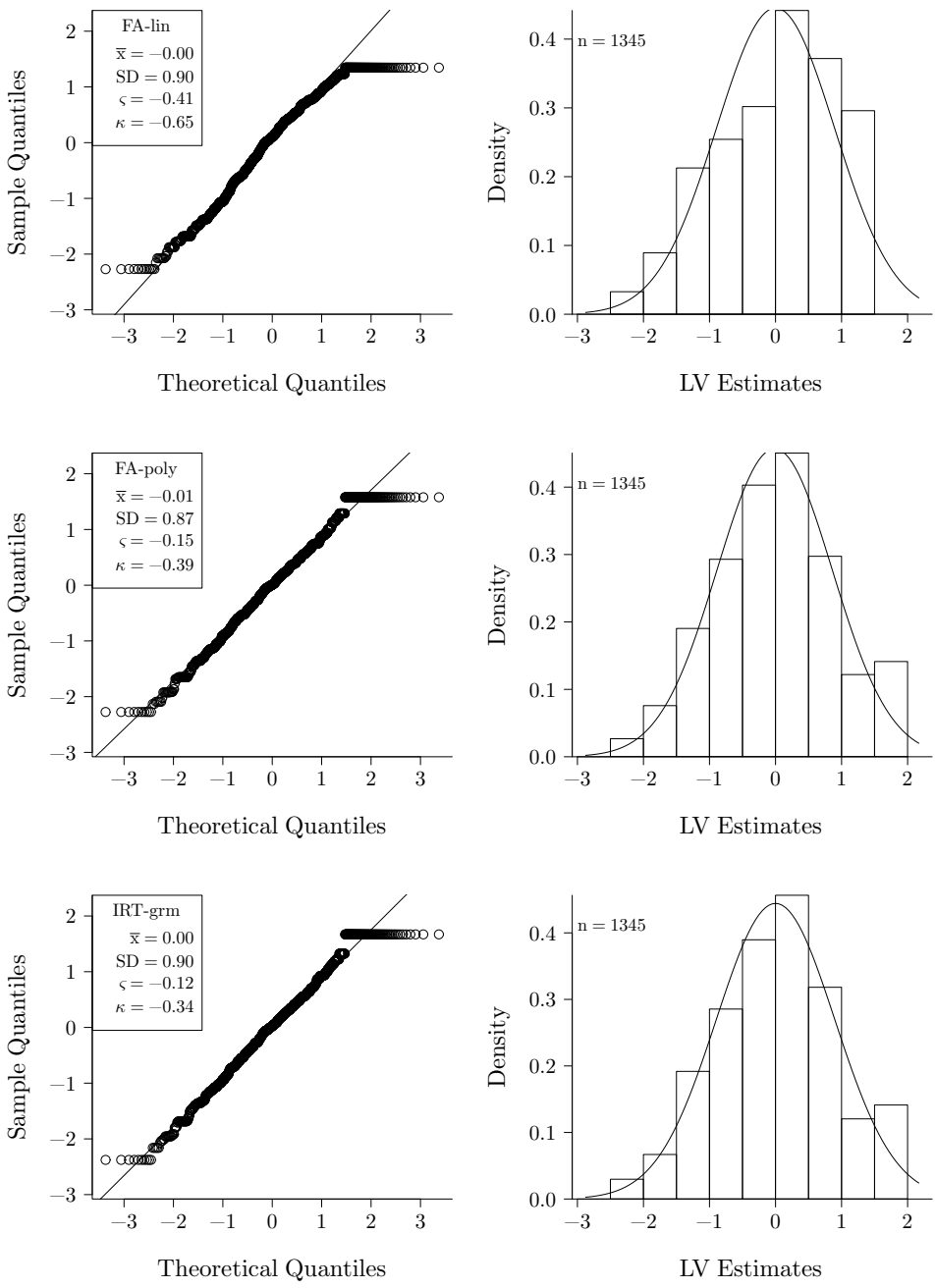



*Figure 7.19.* LV score distribution for total RASJS-*Reactive jealousy*. $n = 1345$.

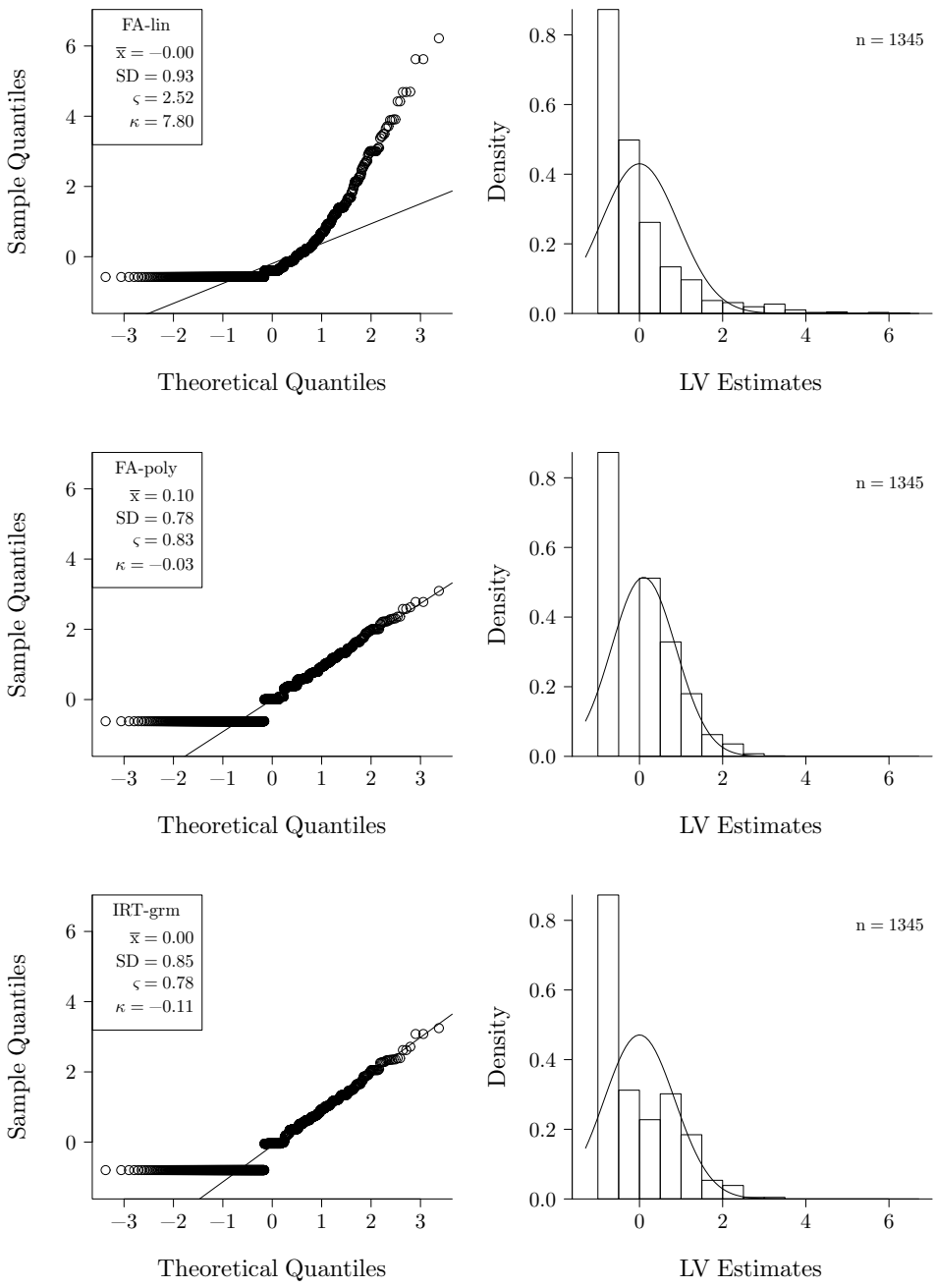



*Figure 7.20.* LV score distribution for total RASJS-*Anxious jealousy*. $n = 1345$.

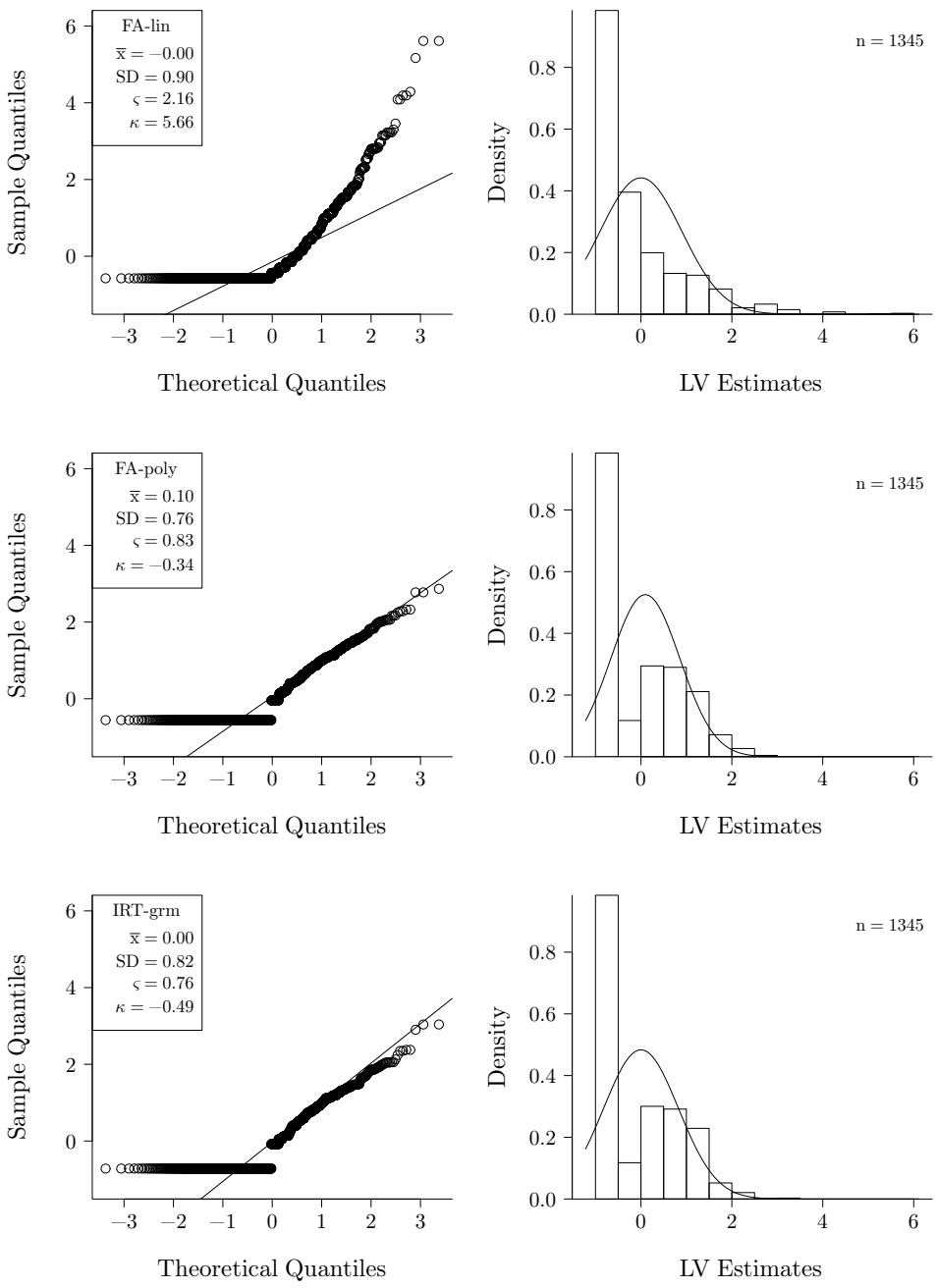



*Figure 7.21.* LV score distribution for total RASJS-*Possessive jealousy*. $n = 1345$.

### Parameter and Standard Error Estimation

Table 7.14 provides the estimated loading parameters per subscale ($\hat{\lambda}_{sub}$) as well as for all items taken as a single dimension ($\hat{\lambda}_{all}$).

We first discuss the results of all items taken together. Recall that in the previous subsection, we inferred the higher level LV *Jealousy* to be normally distributed with mostly right-skewed items. For that case, we expect both FA-poly and IRT-grm to produce accurate parameter estimates. For the first ten items, the FA-poly loading estimates are all larger than the IRT-grm estimates. For the final five items, this pattern is reversed. Based on our simulation study, we would expect FA-poly and IRT-grm loading paramater estimates to be similar, with FA-poly producing slightly larger estimates than IRT-grm, as a result of the skewness of the items (cf. Chapter 6, p. 152) and the size of the loading values (cf. Chapter 5, p. 113). The reason why IRT-grm estimates are larger than those of FA-poly for the final five items is not clear, although the differences are small.

FA-lin loading parameter estimates are all smaller than FA-poly estimates, which corresponds to our simulation study results. Compared to IRT-grm, FA-lin parameter estimates are mostly smaller, except for Items *a1.attractive*, *a3.sex.int*, and *a5.leave*. The IRT-grm parameter estimates of these three items are all rather small.

Standard errors are smallest for FA-poly, followed by FA-lin, and largest for IRT-grm. These differences are most pronounced for the subscales *Reactive* and *Anxious*. Based on our simulation results, we would expect standard errors to be estimated most accurately by IRT-grm, and underestimated by FA-lin and FA-poly.

*Table 7.14.* Loading parameter and standard error estimates for the RASJS subscales *Reactive*, *Anxious*, and *Possessive* jealousy per subscale and for all items together. $n = 1345$.

| | FA-lin | | FA-poly | | IRT-grm | |
|---|---|---|---|---|---|---|
| | $\hat{\lambda}_{sub}$ $(\hat{se})$ | $\hat{\lambda}_{all}$ $(\hat{se})$ | $\hat{\lambda}_{sub}$ $(\hat{se})$ | $\hat{\lambda}_{all}$ $(\hat{se})$ | $\hat{\lambda}_{sub}$ $(\hat{se})$ | $\hat{\lambda}_{all}$ $(\hat{se})$ |
| r1.flirt | 0.696 (0.019) | 0.330 (0.028) | 0.744 (0.017) | 0.574 (0.021) | 0.743 (0.021) | 0.498 (0.040) |
| r2.discuss | 0.572 (0.022) | 0.268 (0.029) | 0.619 (0.021) | 0.450 (0.024) | 0.607 (0.025) | 0.428 (0.035) |
| r3.sex.contact | 0.454 (0.025) | 0.174 (0.030) | 0.767 (0.028) | 0.604 (0.032) | 0.762 (0.027) | 0.530 (0.054) |
| r4.dance | 0.789 (0.017) | 0.312 (0.029) | 0.835 (0.015) | 0.606 (0.020) | 0.830 (0.018) | 0.497 (0.044) |
| r5.kiss | 0.646 (0.020) | 0.287 (0.029) | 0.694 (0.018) | 0.507 (0.022) | 0.694 (0.024) | 0.433 (0.039) |
| a1.attractive | 0.660 (0.018) | 0.625 (0.021) | 0.802 (0.016) | 0.691 (0.015) | 0.808 (0.019) | 0.550 (0.050) |
| a2.sex.rel | 0.782 (0.014) | 0.590 (0.024) | 0.883 (0.015) | 0.805 (0.018) | 0.875 (0.021) | 0.601 (0.067) |
| a3.sex.int | 0.831 (0.013) | 0.605 (0.024) | 0.906 (0.012) | 0.803 (0.016) | 0.896 (0.018) | 0.562 (0.070) |
| a4.general | 0.688 (0.017) | 0.685 (0.019) | 0.786 (0.019) | 0.736 (0.018) | 0.790 (0.025) | 0.714 (0.048) |
| a5.leave | 0.634 (0.020) | 0.554 (0.024) | 0.792 (0.015) | 0.671 (0.018) | 0.807 (0.024) | 0.506 (0.067) |
| p1.meet | 0.764 (0.018) | 0.622 (0.022) | 0.871 (0.016) | 0.810 (0.017) | 0.904 (0.020) | 0.869 (0.024) |
| p2.friend | 0.713 (0.019) | 0.538 (0.024) | 0.826 (0.019) | 0.738 (0.020) | 0.854 (0.025) | 0.800 (0.031) |
| p3.look | 0.562 (0.022) | 0.471 (0.025) | 0.721 (0.025) | 0.655 (0.024) | 0.772 (0.028) | 0.747 (0.030) |
| p4.possessive | 0.563 (0.024) | 0.520 (0.024) | 0.747 (0.020) | 0.679 (0.018) | 0.721 (0.027) | 0.701 (0.025) |
| p5.own.way | 0.707 (0.020) | 0.625 (0.022) | 0.855 (0.015) | 0.770 (0.016) | 0.832 (0.020) | 0.806 (0.021) |

Interestingly, the analyses per subscale result in FA-poly and IRT-grm loading parameter estimates that are much more similar than for the total RASJS scale. This might be caused by the larger true loading values for the subscales which are a result of the narrower concept each LV then represents. As we saw in Chapter 5, smaller population loading parameter values result in more differentiation both within and between estimation methods.

FA-lin subscale parameter estimates are most different for *Reactive jealousy*, compared to the total scale results. This is most likely caused by the fact that these items' distributional shapes vary from uniform to left-skewed, whereas the other ten items are all right-skewed. In the total scale analysis, this presumably resulted in a severe parameter underestimation. Because in the subscale analyses the item distributions are more homogeneous, FA-lin parameter estimation is supposed to be less biased then and more in line with FA-poly and IRT-grm.

Standard errors as estimated by the three models are also more similar for the subscale analyses, although the pattern of results with FA-poly standard errors being the smallest, followed by FA-lin, and IRT-grm standard errors being the largest, remains the same compared to the analysis of the total scale. As the total LV is assumed to be normal and the subscale LVs are thought to be skewed, these results are not in line with our Monte Carlo study results, where we found standard error estimation to be more divergent between the estimation models for a skewed LV than for a normal LV combined with skewed item distributions. The reason for these differences is not clear. Perhaps the small number of items in the subscales is a factor of importance, or the severity of the item and LV skewness, or the large sample size, as compared to our Monte Carlo study.

Threshold parameter and standard error estimates for the total RASJS are presented in Table F.5 of Appendix F. Overall, the FA-poly and IRT-grm threshold parameter and standard error estimates are quite similar.

### Model Fit

Estimated fit indices for the total RASJS are presented in Table 7.15. The upper panel provides the indices calculated in MPLUS. In the lower panel some additional fit indices are listed.

We first notice the large $\chi^2$ values produced by MPLUS for both FA-lin and FA-poly. For IRT-grm no $\chi^2$ could be computed due to the large item cross table (containing $5^{15}$ cells). The corresponding RMSEA values of around 0.17 are not indicative of a good model fit to the data. The CFI and TLI are also quite low, and more so for FA-lin than for FA-poly.

Turning to the additional, non-MPLUS indices, we note the $\chi^2_{YB}$ is much smaller than the ordinary $\chi^2$, but still indicative of a lack of fit, when compared to the degrees of freedom. The RMSEA values of around 0.06 imply an acceptable fit though. The SRMR values, however, are too large when compared to the common criterion value of 0.08.

*Table 7.15.* Model fit results for RASJS. $n = 1345$.

| Fit statistic | FA-lin | FA-poly | IRT-grm |
|---|---|---|---|
| $\chi^2_{mplus}$ | 3448.768 | 3881.891 | |
| df | 90 | 90 | |
| RMSEA | 0.167 | 0.177 | |
| CFI | 0.524 | 0.744 | |
| TLI | 0.444 | 0.701 | |
| $\chi^2_{YB}$ | 624.784 | 246.542 | 253.343 |
| df | 90 | 45 | 45 |
| RMSEA | 0.066 | 0.058 | 0.059 |
| SRMR | 0.137 | 0.136 | 0.137 |

Judging from these fit indices, the unidimensional model should be rejected for each estimation model.

Fit results for the subscale analyses are provided in Tables F.6 to F.8 of Appendix F. As expected, the model fit is notably better for the subscales.

#### Analysis by IRT-mok

The nonparametric IRT-mok model was applied to the subscales as well as to the total RASJS. In Table 7.16 estimated Loevinger's item and scale $H$ coefficients and corresponding standard errors are given for the subscales ($\hat{H}_{sub}$) and the total scale ($\hat{H}_{all}$).

For the total RASJS, most $\hat{H}_i$ values are acceptable if the lower bound criterion of 0.30 were to be applied, except for Items *a1.attractive*, *a3.sex.int*, and *a5.leave*. Notice that these items also drew our attention in the parametric analyses, because their IRT-grm loading parameter estimates were notably small.

From their item plots — only shown for *a5.leave* in Figure 7.22 — it is apparent that these items' item response functions (IRFs) are very flat, indicating a weak association between the item score and the scale score. The monotonicity assumption seems to be violated only minimally.

For the subscale analyses, $\hat{H}$ values are distinctly larger, indicating an increased level of homogeneity, which is, again, the result of the subscale items respresenting narrower concepts than the concept represented by the set of items taken as a whole.

### 7.4.3 Discussion

FA-lin, FA-poly, IRT-grm, and IRT-mok were applied to the RASJS in two separate ways: per subscale and for all items taken together as a single dimension.

Loading parameter estimates and Loevinger's $H$ values were larger in the subscale analyses than in the single-dimension analysis. This illustrates the fact that broadening the concept of a LV by adding items that tap various subconcepts results in weaker associations between items.

*Table 7.16.* IRT-mok results for the RASJS subscales *Reactive*, *Anxious*, and *Possessive* per subscale and for all items together. $n = 1345$.

| Item | $\hat{H}_{sub}$ | $\hat{se}(\hat{H}_{sub})$ | $\hat{H}_{all}$ | $\hat{se}(\hat{H}_{all})$ |
|---|---|---|---|---|
| r1.flirt | 0.501 | 0.018 | 0.376 | 0.017 |
| r2.discuss | 0.440 | 0.020 | 0.304 | 0.017 |
| r3.sex.contact | 0.641 | 0.029 | 0.589 | 0.025 |
| r4.dance | 0.540 | 0.016 | 0.418 | 0.017 |
| r5.kiss | 0.479 | 0.019 | 0.330 | 0.018 |
| $\hat{H}_{scale}$ *Reactive* | 0.498 | 0.016 | | |
| a1.attractive | 0.603 | 0.026 | 0.289 | 0.018 |
| a2.sex.rel | 0.608 | 0.028 | 0.306 | 0.026 |
| a3.sex.int | 0.621 | 0.022 | 0.278 | 0.023 |
| a4.general | 0.567 | 0.030 | 0.338 | 0.022 |
| a5.leave | 0.575 | 0.031 | 0.235 | 0.023 |
| $\hat{H}_{scale}$ *Anxious* | 0.595 | 0.022 | | |
| p1.meet | 0.515 | 0.025 | 0.396 | 0.018 |
| p2.friend | 0.470 | 0.028 | 0.347 | 0.019 |
| p3.look | 0.413 | 0.031 | 0.343 | 0.021 |
| p4.possessive | 0.469 | 0.026 | 0.335 | 0.017 |
| p5.own.way | 0.553 | 0.020 | 0.376 | 0.017 |
| $\hat{H}_{scale}$ *Possessive* | 0.487 | 0.019 | | |
| $\hat{H}_{scale}$ All | | | 0.342 | 0.012 |

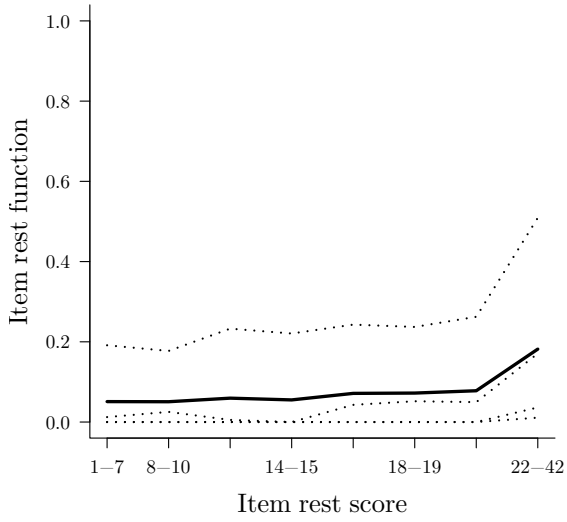



*Figure 7.22.* Item response plot for *a5.leave* based on IRT-mok analysis. The dashed lines represent the item step response functions $P(X_i \geq c)$ for $c = 0, \ldots, 3$. The solid line depicts the mean response function. $n = 1345$.

Taking the items together as a single dimension resulted in a weak scale, from which a small number of items might be considered for removal. Fit statistics were also indicative of a lack of model fit. Analyzing each subscale separately improved the results.

Differences between the parametric models were greater for the single-dimension analysis than for the subscale analyses. FA-poly and IRT-grm parameters were similar for the subscale analyses. Item loadings were all strong (around 0.8) in that case. In the single-dimension analysis, where item loadings diverged more to range between 0.507 and 0.810 for FA-poly, and between 0.428 and 0.869 for IRT-grm, differences between the estimation models were a bit larger.

In all cases, the differences between FA-lin and FA-poly/IRT-grm parameter estimates were the largest. Taking into account our simulation study results, we presume FA-lin loading parameters to be underestimated considerably.

Results from the nonparametric IRT-mok analysis gave a similar general impression, with larger scaling coefficients for the subscale analyses and a weak scale in the single-dimension analysis. In the single-dimension analysis, the items from the second subscale, *Anxious jealousy*, had the smallest $H$ values, which resembled the IRT-grm results.

From a theoretical point of view, the RASJS should be considered a three-dimensional scale. The results from applying the four scaling models support this. The single-dimension analysis demonstrates that taking together related subconcepts in a unidimensional scale can result in smaller item loadings — or scaling coefficients — and a weaker though still acceptable scale. Just as we saw for the DBIQ, taking together the subscales in a single super-scale smooths out the differences of the sublevel components. The resulting normal LV distribution of the total scale could be considered advantageous for further analyses involving the LV scores.

Whether a scale such as the RASJS should be employed and interpreted unidimensionally or multidimensionally depends on the applied researcher's objectives: If one wants to obtain a measurement of a single, broad, high-level concept, the unidimensional scale will do; if one rather needs a more detailed measurement of the multiple aspects that concept entails, the multidimensional scale is most useful. In summary, it depends on the researcher's theoretical considerations regarding LVs.

Rather than a true multidimensional analysis, we carried out three unidimensional analyses. Of course, one could also employ a true multidimensional analysis, as we did in the first application regarding the DBIQ. In fact, we applied a multidimensional analysis with results similar to the three unidimensional analyses. Contrary to our results regarding the DBIQ, we did not find large differences in IRT-grm loading parameter estimates between the unidimensional and multidimensional analyses. This is related to the fact that the estimated multivariate LV correlations were very similar for the three parametric models, in constrast to the multivariate DBIQ LV correlations which differed between IRT-grm and the FA models.

## 7.5 Involvement in Neighbourhood Community Scale

As part of her dissertation research, Frieling (2008) developed a scale to measure the level of social cohesion in a neighborhood. The Involvement in Neighbourhood Community Scale (INCS), a five-point Likert scale, was used to investigate how the communal strength — or social cohesion — of neighborhoods can be monitored and enhanced. By means of an IRT-mok analysis Frieling eventually constructed a short version of her scale, here to be denoted as INCS-*s*, consisting of seven items, and subject to our investigations.

In this section, the data of Frieling's (2008) second sample are used, consisting of $n = 255$ respondents (122 men, 133 women). These data are from a community sample and were collected by a data collection company. The 1.9% missing data, found to be missing at random, were imputed using a two-way imputation method (Bernaards & Sijtsma, 2000). That imputed data set is used in the upcoming analyses.

We first provide descriptive statistics of the sample, next turn to the results of applying the four scaling models to these data, and finally discuss those results.

### 7.5.1 Descriptive Statistics

In Table 7.17 item descriptions and means and standard deviations are provided for the INCS-*s*. From the descending item means, it can be observed that the items were ordered from easy to difficult.'Easy' and 'difficult' are here to be interpreted as requiring, respectively, little and much from the LV *involvement in neighborhood community*.

Item and sum score distributions of the INCS-*s* are presented in Figure 7.23. The distributional forms of the items are quite diverse, including left-skewed, right-skewed, bimodal, and uniform shapes. This is not so surprising, considering that one of the criteria employed in Frieling's selection of the short scale items was diversity in item means. Nonetheless is it remarkable that not even the items in the middle of the scale resemble a normal distribution at all. The sum score distribution is mildly right-skewed and platykurtotic ($\varsigma = 0.13$, $\kappa = -0.75$).

*Table 7.17.* Items of the Involvement in Neighbourhood Community Scale—Short version.

| Abbreviation | Item Description | Mean[a] | SD |
|---|---|---|---|
| 1.talk | Within the last half year, how often have you talked to someone living in your neighborhood?[b] | 2.99 | 1.21 |
| 2.property | Is there someone living in your neighborhood who keeps an eye on your property when you are away from home, e.g., by being watchful for burglars or by taking care of your pet(s) or plants?[c] | 2.65 | 1.70 |
| 3.inform | When something important happens in your neighborhood or to one of the residents, is there someone in the neighborhood who lets you know and keeps you up to date?[c] | 2.11 | 1.67 |
| 4.connected | Do you feel connected to the people on your block?[d] | 2.02 | 1.36 |
| 5.support | When there is a sorrowful event or when something drastic happens in your life, is there someone in your neighborhood who supports you?[c] | 1.55 | 1.71 |
| 6.party | Are there neighborhood or street parties or other activities in your neighborhood where all neighborhood residents are invited? If so, how often do you attend these parties or activities?[c] | 0.94 | 1.47 |
| 7.organize | Within the last year, did you cooperate with other neighborhood residents in organizing something for the neighborhood, e.g., organizing a neighborhood or street party or participating in putting a local paper together? If so, how often did you meet with these other neighborhood residents in the last year?[e] | 0.27 | 0.86 |

*Note:* The item descriptions were taken from Frieling (2008, p. 243).
[a] Item means and standard deviations were computed from the sample of size $n = 255$.
[b] Response categories: *"Once a year or less often"* (0), *"Several times a year"* (1), *"Several times a month"* (2), *"Once a week"* (3), *"Several times a week or more often"* (4).
[c] Response categories: *"Hardly ever"* (0), *"Mostly not"* (1), *"Sometimes/Sometimes not"* (2), *"Mostly"* (3), *"Almost always"* (4).
[d] Response categories: *"To hardly anyone"* (0), *"Not to most people"* (1), *"To some I do, to others I don't"* (2), *"To most people"* (3), *"To almost everyone"* (4).
[e] Response categories: *"[Not cooperated]"* (0), *"Met approximately once every half a year or less often"* (1), *"Met approximately once every three months"* (2), *"Met approximately once every two months"* (3), *"Met approximately every month or more often"* (4).

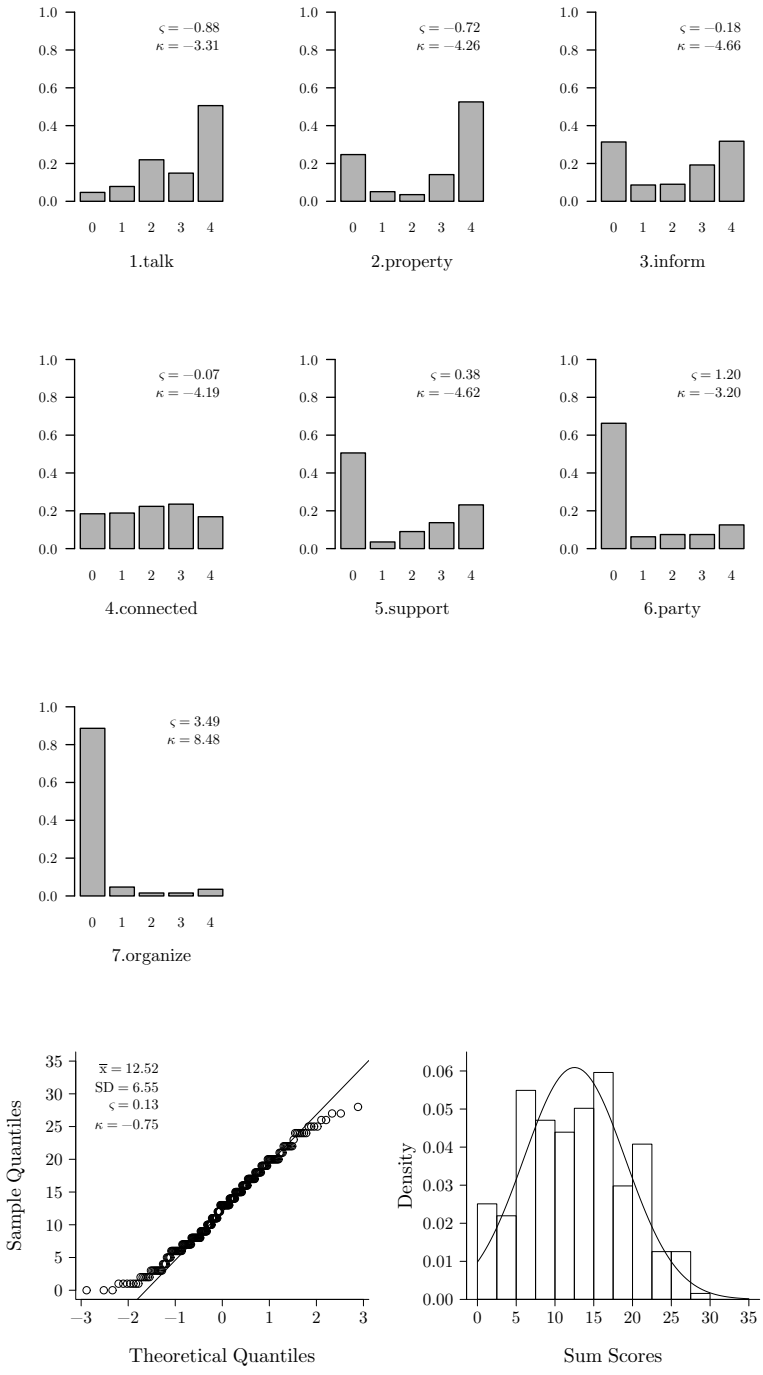



*Figure 7.23.* Item barplots and sum score Q-Q plot and density plot for INCS-*s*. $n = 255$.

### 7.5.2 Results

#### LV Score Estimation

LV score distributions resulting from the FA-lin, FA-poly, and IRT-grm analyses are presented in Figure 7.24. The IRT-mok LV score distribution equals the sum score distribution, and was presented in Figure 7.23. Skewness and kurtosis information is collected in Table 7.18.

Skewness values are rather similar among the parametric models (0.00 to 0.03). IRT-mok stands out a bit with a value of 0.13. The LV distribution may therefore be presumed not to be skewed.

The excess kurtosis of the LV estimates — which can be observed from the "fat" tails of the distribution — is more variable: FA-lin has the largest (absolute) value, followed by IRT-mok, and, finally, the almost equivalent FA-poly and IRT-grm. Since symmetric kurtotic LVs were not part of our simulation study design, we cannot directly conclude what this implies for the shape of the LV distribution. Since the kurtosis is not extreme, this LV most resembles the normal LV of our simulation study. We shall presume this LV to be normal, albeit slightly platykurtic.

*Table 7.18.* Skewness and kurtosis of LV score estimates for INCS-$s$; $n = 255$

| | IRT-mok | FA-lin | FA-poly | IRT-grm |
|---|---|---|---|---|
| Skewness | 0.13 | 0.03 | 0.01 | 0.00 |
| Kurtosis | −0.75 | −0.94 | −0.46 | −0.42 |

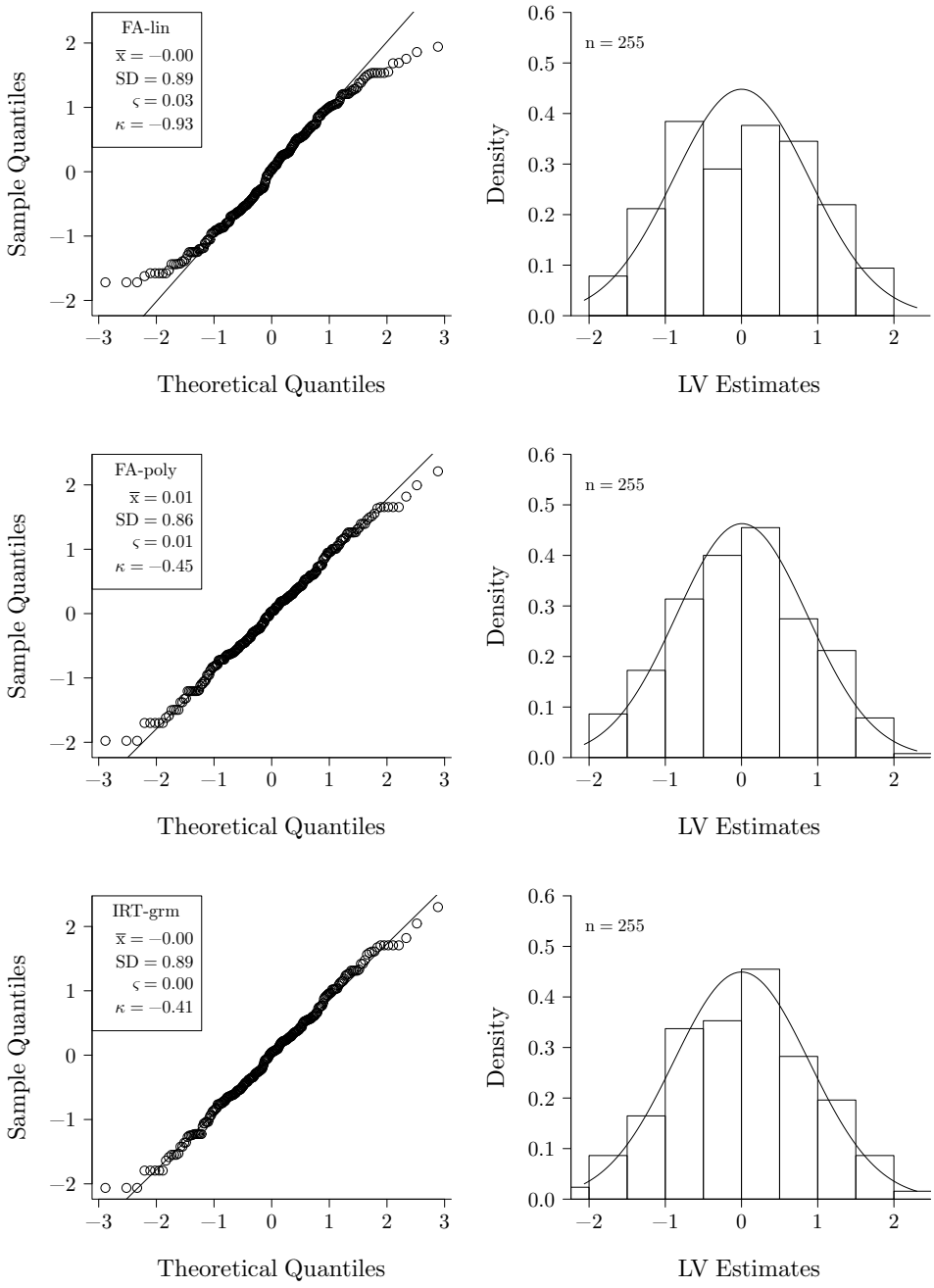



*Figure 7.24.* LV score distribution for INCS-*s*. $n = 255$.

*Table 7.19.* Loading parameter and standard error estimates for INCS-*s*. $n = 255$.

| | FA-lin | | FA-poly | | IRT-grm | |
|---|---|---|---|---|---|---|
| | $\hat{\lambda}$ | $\hat{se}(\hat{\lambda})$ | $\hat{\lambda}$ | $\hat{se}(\hat{\lambda})$ | $\hat{\lambda}$ | $\hat{se}(\hat{\lambda})$ |
| 1.talk | 0.558 | 0.053 | 0.646 | 0.053 | 0.614 | 0.055 |
| 2.property | 0.543 | 0.054 | 0.675 | 0.051 | 0.642 | 0.062 |
| 3.inform | 0.668 | 0.046 | 0.711 | 0.046 | 0.703 | 0.052 |
| 4.connected | 0.667 | 0.046 | 0.675 | 0.047 | 0.657 | 0.053 |
| 5.support | 0.680 | 0.045 | 0.768 | 0.048 | 0.743 | 0.054 |
| 6.party | 0.455 | 0.059 | 0.563 | 0.063 | 0.519 | 0.074 |
| 7.organize | 0.363 | 0.063 | 0.642 | 0.074 | 0.610 | 0.104 |

### Parameter and Standard Error Estimation

Loading parameter estimates are presented in Table 7.19. Loading estimates are comparable for the various models, with FA-poly and IRT-grm estimates being more alike and FA-lin estimates consistently smaller. This difference is largest for skewed items such as Item *7.organize* ($\lambda = 0.36$ for FA-lin versus $\lambda = 0.64$ for FA-poly and $\lambda = 0.61$ for IRT-grm). Presuming the LV to be normally distributed, we expect FA-lin parameters to be underestimated and FA-poly and IRT-grm parameters to be accurately estimated for any item distribution.

Loading standard error estimates are also rather similar for the three models, with FA-lin and FA-poly estimates being more alike, and IRT-grm estimates being larger. Considering our Monte Carlo results, where we found FA-lin and FA-poly loading standard errors to be underestimated — unacceptably for FA-lin and marginally acceptably for FA-poly — in case of skewed items loading on a normal LV, and IRT-grm standard error estimators to be unbiased, we take the IRT-grm standard errors to be most accurate. The most difficult item in the scale, Item *7.organize*, has the largest standard error, which is probably a result of this item's extreme skewness.

Threshold parameters and standard errors as estimated by FA-poly and IRT-grm are presented in Table F.9 of Appendix F. In line with the loading results, threshold parameters and standard errors are similar for FA-poly and IRT-grm.

### Model Fit

Fit statistics for the INCS-*s* are provided in Table 7.20, with MPLUS output in the upper panel and additionally calculated fit indices in the lower panel.

The FA-lin statistics are indicative of a mediocre fit. For FA-poly most indices point towards a good model fit, except for the $\chi^2$, which is more than twice the size of its number of degrees of freedom, and the RMSEA, which is a bit high. For IRT-grm notice, as we did for the DBIQ, the large $\chi^2$ value, which is, however, considerably smaller than its number of degrees of freedom. The SRMR is clearly indicative of a good model fit.

*Table 7.20.* Model fit results for INCS-*s*. $n = 255$.

| Fit statistic | FA-lin | FA-poly | IRT-grm |
|---|---|---|---|
| $\chi^2_{mplus}$ | 44.802 | 33.591 | 2987.778 |
| df | 14 | 14 | 78050 |
| RMSEA | 0.093 | 0.074 | |
| CFI | 0.916 | 0.971 | |
| TLI | 0.873 | 0.956 | |
| $\chi^2_{YB}$ | 31.613 | | |
| df | 14 | | |
| RMSEA | 0.070 | | |
| SRMR | 0.058 | 0.053 | 0.056 |

**Analysis by IRT-mok**

In Table 7.21 the results are presented from applying the nonparametric IRT-mok to the INCS-*s* data. With all items' scaling coefficients larger than 0.40 and a scale $H$ of 0.451, the items make a strong, homogeneous scale.

The variation in item means ensures that the items cover a broad range of the LV scale. This is apparent from the item response plots, two of which are provided in Figures 7.25 and 7.26. From the item response plot of Item *2.property* one can infer that this item mostly distinguishes between respondents scoring on the lower and the middle area of the latent scale, as the ISRFs are least diverging for the highest item rest score group. Item *7.organize* clearly only discriminates between respondents on the higher end of the latent scale, although the discrimination does not appear too large, when considering the still narrow vertical diversion of ISRFs at the highest item rest score group. These two items each contribute uniquely to the scale, making it suitable for a broad population of respondents.

*Table 7.21.* IRT-mok results for INCS-*s*; $n = 255$

| Item | $\hat{H}$ | $\hat{se}(\hat{H})$ |
|---|---|---|
| 1.talk | 0.425 | 0.043 |
| 2.property | 0.427 | 0.046 |
| 3.inform | 0.457 | 0.040 |
| 4.connected | 0.443 | 0.038 |
| 5.support | 0.489 | 0.038 |
| 6.party | 0.404 | 0.051 |
| 7.organize | 0.592 | 0.062 |
| $\hat{H}_{scale}$ | 0.451 | 0.033 |

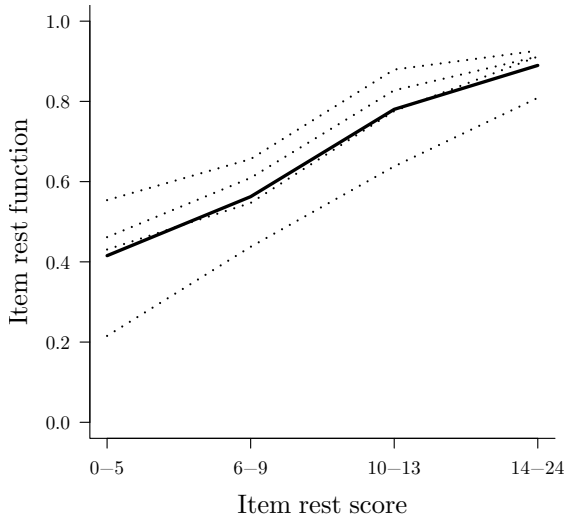



*Figure 7.25.* Item response plot for *2.property* based on IRT-mok analysis. The dashed lines represent the item step response functions $P(X_i \geq c)$ for $c = 0, \ldots, 3$. The solid line depicts the mean response function. $n = 255$.

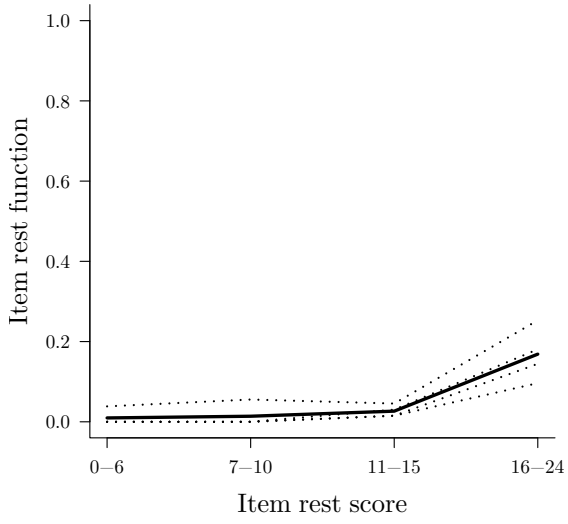



*Figure 7.26.* Item response plot for *7.organize* based on IRT-mok analysis. The dashed lines represent the item step response functions $P(X_i \geq c)$ for $c = 0, \ldots, 3$. The solid line depicts the mean response function. $n = 255$.

### 7.5.3 Discussion

The INCS-*s* was analyzed by means of FA-lin, FA-poly, IRT-grm, and IRT-mok. None of the items of the INCS-*s* were even approximately normally distributed. Nevertheless, all parametric models seemed to perform rather well.

All results indicated that the items are associated strongly enough to form a scale, although, in comparison, FA-lin produced the smallest item loadings and the worst model fit.

As we tentatively interpreted the LV estimation results to be indicative of a normal but slightly platykurtotic LV, FA-poly and IRT-grm parameter estimates were considered to be accurate, whereas FA-lin loading parameters were likely to be underestimated. Loading standard errors were interpreted to be most accurate as estimated by IRT-grm.

Since a symmetrical kurtotic — or perhaps truncated — LV was not incorporated in our Monte Carlo design, these kinds of LVs call for some further research.

## 7.6 Discussion

In this chapter we presented a number of applications of the scaling models under investigation. We thus demonstrated the practice of employing FA of the sample covariance matrix (FA-lin), FA of the estimated polychoric correlation matrix (FA-poly), the graded response IRT model (IRT-grm), and the nonparametric Mokken IRT model (IRT-mok) in three empirical settings, taking into account the findings of the Monte Carlo study laid out in the previous chapters.

#### Summary of the Three Applications

The first application concerned the Dresden Body Image Questionnaire (DBIQ), a scale consisting of 35 items loading on five subdimensions submitted to a sample of 761 respondents. We applied all scaling models separately to the five sub-scales. In addition, we employed the parametric models multidimensionally. Based on our Monte Carlo findings, the distributional shapes of the latent variables (LVs) corresponding to the sub-scales were presumed to be diverse, including left-skewed, normal and right-skewed types. Item distributions were in majority skewed. These characteristics made the FA-lin model least appropriate for the data. Based on the simulation study, FA-poly and IRT-grm were judged to perform well. The nonparametric IRT-mok was also considered suitable for these data. The empirical results of these three models were comparable and designated to be trustworthy. The model chosen to employ would not affect the results or consequences of the scale analysis in an essential way. Using FA-lin could result in an underestimation of the scale strength, but probably not to discarding items that would have been retained by employing FA-poly or IRT-grm.

Furthermore, for FA-lin and FA-poly, multidimensional results much resembled those from the five unidimensional analyses. For IRT-grm, however, some striking

differences were found between unidimensional and multidimensional estimated loading parameters, with deviations as large as 0.49. In accordance with the loading parameter estimation discrepancies, the estimated LV covariances differed between FA-lin/FA-poly and IRT-grm. The reasons for these divergences are not clear and have to be investigated in future research.

The second application provided the opportunity to further explore the consequences of applying an estimation model to a sub- or super-level of a theoretical construct. The Revised Anticipated Sexual Jealousy Scale (RASJS) is a 15-item scale measuring *sexual jealousy* on three sub-scales of 5 items each, and was employed to a sample of 1366 respondents. We applied the scaling models both to the sub-levels (i.e., three unidimensional analyses of the 5-item scales) and to the super-level (i.e., a unidimensional analysis of all 15 items).

Of the three LVs, one was presumed to be left-skewed and two were probably highly right-skewed, whereas the super-level LV was thought to be normal. Item distributions of two of the sub-scales were right-skewed, and varied for the remaining sub-scale (uniform, left-skewed, bimodal/U-shaped). FA-lin was not found to be suitable for these data. The small estimated FA-lin item loadings compared to FA-poly and IRT-grm estimates supported the findings from our simulation study, where FA-lin was found to underestimate loading parameters. For the sub-scale analyses, FA-poly and IRT-grm parameter estimation results were almost identical.

For the super-scale analysis, FA-poly and IRT-grm results diverged slightly more, but not to a great extent. Presumably, the larger differences for the super-level analyses were caused by the smaller true values of the item loadings in this case. In our Monte Carlo study, we found differences in loading parameter estimation between FA-poly and IRT-grm to increase for decreasing population values.

FA-lin estimates differed substantially more from FA-poly and IRT-grm in case of the super-level analysis. Employing FA-lin would result in other conclusions than employing FA-poly or IRT-grm. A number of items would likely be discarded based on the loading parameter estimates resulting from an FA-lin analysis, whereas their relation to the LV would be considered sufficiently strong based on FA-poly or IRT-grm.

For both the parametric and the nonparametric models, the sub-scale analyses led to better results, with higher item loadings and a better model fit. The super-scale analysis demonstrated that taking together interrelated LVs at a higher conceptual level can result in smaller item loadings and a weaker though still acceptable scale.

In the third application, a short version of the Involvement in Neighbourhood Community Scale (INCS) was reanalyzed, by means of a sample of size $n = 255$. Possibly resulting from the fact that this scale was constructed with the aid of employing the IRT-mok model, all seven items' distributional shapes were distinctly nonnormal. From the LV estimation results we derived the LV to be slightly platykurtic but not skewed. Such a LV distribution was not part of our Monte Carlo design, and thus the results were not easily matched to the findings from that study. Future research is needed incorporating nonskewed kurtotic LVs.

Compared to FA-poly and IRT-grm, we found FA-lin loading estimates to be small and model fit to be poor, as caused by the nonnormality of the items. In practice, selecting FA-lin as the model of choice would result in an underestimation of the scale strength. FA-poly, IRT-grm, and IRT-mok all produced similar results, and any one could be elected as the model for analyzing these data.

**Conclusion**

For all empirical analyses in this chapter, we note that in case of nonnormal item distributions, FA-lin loading parameter estimates are small compared to those of FA-poly and IRT-grm. This is in accordance with the findings from our simulation study, where we encountered considerable underestimation.

In general, FA-poly and IRT-grm results were similar but not equal. Differences were not large enough to lead to different conclusions, though. From our simulation study results, we have a slight preference for IRT-grm in case of strong factor loadings and a skewed LV. With smaller loading values, FA-poly parameter estimation seems to be superior to IRT-grm, at least for normal LVs and items. Standard errors are, generally, estimated more accurately by IRT-grm. Furthermore, fit indices, at least those produced by MPLUS, are more abundant for FA-poly than for IRT-grm and potentially provide more useful information when assessing model fit.

The results of the IRT-mok analyses were found to resemble those of the parametric FA-poly and IRT-grm models, with the scaling coefficients being comparable to the loading estimates. The advantage of FA-poly and IRT-grm over IRT-mok lies mostly in the more sophisticated LV estimation of the former two in case of discongruent LV and item distributions. Since thresholds are taken into account in FA-poly and IRT-grm LV estimation, the LV distribution is estimated more accurately in such cases. IRT-mok merely leads to an ordering of respondents on the LV, leaving the LV distribution unidentified.

We therefore advise the application of both FA-poly and IRT-grm, after a thorough examination of the item and sum score distributions, as was employed in the present chapter. Results of both models should be compared and the interpretation of the results depends on the characteristics of the data and the parameter estimates. For example, in case of medium to small factor loadings and diverging parameter estimation results for FA-poly and IRT-grm, one should put more trust in the FA-poly estimates. However, in case of a skewed LV and skewed item variables, IRT-grm estimates should be preferred. Our simulation study demonstrated that standard error estimators are more accurate for IRT-grm than for FA-poly, with FA-poly generally underestimating the standard errors. This finding was supported by the fact that in each application FA-poly estimated standard errors were smaller than those of IRT-grm. IRT-grm standard errors are thus to be favored. Fit of the model can be ascertained using the FA-poly fit statistics.

In our Monte Carlo study we found FA-poly and IRT-grm LV estimates to be more normally distributed than their true counterparts: Population skewness was recovered, but underestimated. So, as a practical rule, when LV skewness is found using

FA-poly or IRT-grm, the true LV is likely to be even more skewed, and a model with an even more skewed LV might be more appropriate. Notably, even in case of left-skewed items loading on a right-skewed LV, the correct direction of LV skewness was recovered, which was never accomplished by FA-lin or IRT-mok. This finding from our simulation study was empirically supported here.

# Chapter 8

# Discussion

## 8.1 Summary

In this dissertation, the differences and similarities between factor analysis (FA) and item response theory (IRT) were investigated with respect to the stability and sensitivity, i.e., the robustness, of their results to violations of distributional assumptions. In the introduction to this dissertation, we presented the general practice of employing FA and IRT modeling in the process of constructing and evaluating a scale, the central research questions, and the approach taken to resolve those questions.

After a brief discussion of the concept of a latent variable (LV), the two approaches to scale analysis under investigation, FA and IRT, were introduced in Chapter 1. From two traditions that developed independently, scaling models evolved that bear a number of similarities. Mathematically, some FA and IRT models are even equivalent. FA and IRT also have their respective traditions in estimation methods being applied, with FA predominantly linked to limited-information (LI) and IRT connected to full-information (FI) methods.

In Chapter 2 we turned to the practice of scale analysis, performing a review of journal articles in which FA and IRT models were employed in scale construction and evaluation. We found that FA was applied far more often than IRT and justifications for using either method were often lacking, forcing us to make some educated guesses about researchers' reasons for choosing a scaling model for their analyses. Expectations about the dimensionality of the scale could be a motive to apply FA instead of IRT. Perceived lack of easily accessible (multidimensional) IRT software could be another reason for preferring FA over IRT.

We found that model assumptions, such as those regarding the distribution of the data, were often not investigated, or at least left unreported. This is troublesome because often a linear FA model was applied to ordered categorical data, which has been found to be acceptable only if the item variables are approximately normally distributed (e.g., Boomsma, 1983; Coenders et al., 1997; Flora & Curran, 2004;

Hoogland, 1999; Jöreskog & Moustaki, 2001; Moustaki et al., 2004; B. O. Muthén & Kaplan, 1985, 1992). These findings gave rise to questions about the consequences of applying scaling models whose assumptions are violated, supporting the practical motivation for this research.

In Chapter 3 we presented an overview of previous simulation research focused on FA of ordered categorical data, two-parameter IRT models, or both. The studies were summarized and chronologically listed in two comprehensive tables (see pp. 67 and 68). Based on the reviewed literature we formulated general expectations regarding the robustness questions for our own simulation study, and justified the explanatory factors of interest in our research design: (a) LV distribution, (b) item response distribution, (c) scale strength, and (d) sample size. The data configurations resulting from combining these factors were to be analyzed by applying each of four selected scaling models: (a) FA of the sample covariance matrix with maximum likelihood estimation (e.g., Jöreskog, 1967; FA-lin-ML); (b) FA of the estimated polychoric correlation matrix (Olsson, 1979) using mean-and-variance adjusted weighted least squares (B. O. Muthén, 1984; FA-poly-WLSMV); (c) the graded response IRT model (Samejima, 1969) with robust ML (L. K. Muthén & Muthén, 1998–2010, p. 533; IRT-grm-MLR); (d) The nonparametric Mokken IRT model (Mokken, 1971) extended to polytomous items (I. W. Molenaar, 1982). FA-lin-ML is included as the standard practice, although, by definition, the FA-lin model does not hold for discrete item variables. FA-poly and IRT-grm are both included, even though they have been shown to be theoretically equivalent (Takane & De Leeuw, 1987), because the models are typically estimated using different estimation methods, and the theoretical equivalence only holds under the assumption of normality.

In Chapter 4 the setup of our simulation study was presented, explaining the data generation process and addressing the performance variables and criteria to be applied in the evaluation of results. We also specified our expectations with regard to the performance variables of the simulation study, heavily leaning on the literature discussed in Chapter 3. We divided our design into two parts: Normal data configurations — with approximately normal items loading on a normal LV — were addressed in Chapter 5, providing a benckmark for the nonnormal data configurations presented in Chapter 6.

We expected little differences in the performance of the four models for normal data conditions, apart from a small but consistent negative bias for FA-lin loading parameter estimators known from previous studies. For LV and/or item distributions deviating from normality, we expected larger differences between model performances. FA-lin was expected to be affected most, with severely biased parameter estimators, except when item thresholds were evenly spaced, i.e., in case of equally skewed LV and item variables. FA-poly parameter estimation was expected to be more adversely affected by nonnormality of the LV distribution than by nonnormality of item distributions. As for the parametric models, IRT-grm was expected to perform relatively well in conditions of nonnormality. Because of its nonparametric character, IRT-mok was not expected to be affected much by nonnormality of the data.

In Chapter 5 results were presented of applying the four estimation models to generated samples of normal data. Data were configured as unidimensional scales consisting of 12 five-category items. Items were either all associated strongly to the LV, or of mixed strength, i.e., four strong, four medium, and four weak item-LV associations. Sample size was either small ($n = 200$) or medium ($n = 600$).

In Table 5.11 (p. 139) a summary of the results regarding the parametric models was given, referring to our prior expectations. Most of our expectations with regard to the normal data configurations were supported. Two exceptions were those regarding FA-lin standard error and LV score estimators, which were found to be more accurate than expected. Both FA-poly and IRT-grm performed well with respect to all performance criteria, which was in line with our expectations.

With regard to the nonparametric IRT-mok, the scalability coefficient Loevinger's $H$ was found to be consistently positively biased to the small extent of 5%, decreasing to zero for very large sample sizes (see Section 5.2.5). The bias was presumably caused by the fact that all population item means were set equal to zero, complicating the computation of $H$, which is based on the item ordering within a sample. Standard error estimators for $H$, recently made available by Kuijpers et al. (2013), were found to be unbiased.

In Chapter 6 a Monte Carlo study regarding nonnormality was carried out. Data configurations differed with respect to the included item distributions (normal, skewed, or bimodal), LV distribution (normal or skew-normal), and sample size (small or medium), while keeping the item-LV associations constant and strong.

In Table 6.11 (p. 188) a summary of the results for the parametric models was presented with reference to our expectations brought forth in Chapter 4. As was apparent from the left part of that table, FA-poly and IRT-grm performed well with regard to every performance variable included in our study in case of a normal LV distribution, regardless of the item distribution, outperforming FA-lin. In case of a skew-normal LV, we found the performance of all parametric models to deteriorate, most notably when combined with skewed item variables. IRT-grm performed best in such conditions, outperforming FA-poly with regard to parameter and standard error estimation.

With regard to IRT-mok, we concluded that the nonnormal LV and item distributions did not pose any estimation problems. On the contrary, scales with heterogeneously shaped item distributions led to increased $H$ values that were estimated more accurately than homogeneous scales. Corresponding standard error estimators were found to be unbiased in all conditions.

The results from Chapters 5 and 6 clearly demonstrated that LV estimates based on estimated model parameters are more informative about a respondent's LV score than unweighted sum scores, especially in case of incongruous LV and item distributions (e.g., skewed items loading on a normal LV), or scales with item loadings of different size.

In Chapter 7 we returned to the practice of scale analysis. We presented a number of applications of the scaling models under investigation, demonstrating the practice

of employing FA-lin, FA-poly, IRT-grm, and IRT-mok in three empirical settings. Applying the four scaling models, we were able to make inferences about the LV distribution based on the findings of our Monte Carlo study. Even though the population parameters were — naturally — unknown in these empirical settings, it was possible to relate most results to findings from our simulation study, thus aiding in the interpretation of the estimation results. We found that applying FA-lin would generally result in an underestimation of the item-LV association, and could in some cases misguide the scale analyst into discarding items that would have been considered sufficiently strongly related to the LV employing FA-poly, IRT-grm, or IRT-mok.

## 8.2 Guidelines

In this section we provide some guidelines for applied researchers when employing scale analysis.

#### Take a Large Enough Sample Size

Preceding the collection of data, one should contemplate the purpose the data will be collected for and the analyses required to realize one's ambitions. These elements determine the required sample size. In our study, all parametric models produced rather unprecise estimators, i.e., the variance of the estimators was relatively large, in case of $n = 200$. IRT-mok's $H$ value estimation was also much more precise for the sample size of 600. Therefore, employing a sample size as small $n = 200$ is not recommended, as the reliability and hence usefulness of a parameter or standard error estimate is questionable then.

#### Use Available Substantive Knowledge

In our literature review of Chapter 2 we found many researchers applying exploratory FA to items written to be indicative of a certain trait or LV, even to validate a scale, thus discarding their substantive knowledge about the items. We recommend that applied researchers use such knowledge to their advantage, and employ a confirmatory model testing the hypothesized structure of the data (cf. I. W. Molenaar, 1988). Although we did not investigate this topic any further in our simulation study, we advise that item content and the interpretability of LVs are taken into account in the evaluation of a model, in addition to the estimates of model parameters such as loadings and corresponding standard errors as well as model fit indices.

With regard to IRT-mok, the fact that the estimated scalability coefficient $H$ was found to depend on both the sample size and the dispersion of item locations also stresses the importance of not applying criteria for item retention such as $H > 0.3$ too strictly. Item and LV content should always be taken into careful consideration in scale analysis.

### Inspect the Sample Data

Although this is certainly not new, researchers are — once again — advised to inspect their data. It is new that this may help them to select an appropriate scaling model. They should examine the item variables and the scale's sum scores using histograms or other graphical tools, as was done in Chapter 7. As we have argued, the distributional shape of the sum scores determines to a large extent which model is likely to produce the best estimation results. Furthermore, the distributional properties of the items can be taken into account when deciding on the removal of specific items from a scale. As the greatest potential problem for an IRT-mok analysis might be a scale composed of items with equal item means, it is of importance to examine the dispersion of the item means before applying the IRT-mok model. It is furthermore recommended to (briefly) report the results of such explorations.

### Choose a Model

Model choice depends on both the properties of the data and the purpose of the scale analysis, as explained below.

When items are ordered categorical, the use of FA-lin with standard ML estimation is advised against, because FA-lin loading parameters are likely to be underestimated and, in case of item and/or LV skewness, FA-lin standard error estimators are underestimated as well and model fit cannot reliably be estimated by means of the root mean squared error of approximation (RMSEA). Therefore, applying FA-lin may result in an underestimation of the item-LV relation, possibly with an overestimated precision. This, in its turn, may misguide the researcher into eliminating informative items, ultimately resulting in a less reliable scale.

IRT-mok should not be employed if one requires interval level LV scores, because IRT-mok merely results in an *ordering* of respondents on the LV. In case the LV scores are used in subsequent analyses, the measurement level required in these analyses determines whether IRT-mok suffices. For example, interval level is required when the LV scores are to be used in an "ordinary" regression analysis, whereas ordinal measurement level suffices when the LV scores are to be associated with another ordinal variable by employing Kendall's $\tau$, for instance.

Which model ought to be employed can be further determined by examining the distributional shape of the sum scores. If it is normal, the LV may be presumed to be normal too. In that case, FA-poly, IRT-grm, or IRT-mok can all be applied, expecting accurate results that are precise if the sample size is large enough ($n = 200$ could be too small, but $n = 600$ suffices). If the model-estimated LV distribution deviates from normality, the true LV distribution is presumably also nonnormal. In that case, it is advised to employ either IRT-mok or IRT-grm. Since parameter and standard error estimation was found to be more accurate for IRT-grm than for FA-poly, IRT-grm is preferred over FA-poly. However, FA-poly could be used in addition to IRT-grm, because of its useful fit statistics.

As long as the limitation of ordinal LV scores is not decisive for model choice, IRT-mok can be employed, also for nonnormal LVs. Items with similar sample means are least favorable for IRT-mok. In addition, as we found the estimation of the scalability coefficient $H$ to be rather unprecise for the sample size of $n = 200$ as compared to $n = 600$, we would not necessarily advise the use of IRT-mok over either FA-poly or IRT-grm in case of a small sample size.

#### Assess Model Fit

FA-lin model fit indices are unreliable in case of skewed item variables. For any of the LV and item distributions, the $\chi^2_{YB}$ fit statistic, the RMSEA based on it, and the standardized root mean residuals (SRMR) can be used for assessing model fit for FA-poly and IRT-grm, as they showed acceptable performance. Since these statistics are generally not available for IRT-grm — we implemented them in R, based on Maydeu-Olivares et al. (2011) — they should be noted as useful additions to the few model fit statistics at hand for IRT-grm. However, the $\chi^2_{YB}$ and the RMSEA based on it are only of use if the number of items is at least twice the number of categories, i.e., when the model is identified from the sample covariance matrix.

When using MPLUS one could employ both FA-poly and IRT-grm, using FA-poly to assess model fit. In case the fit statistics indicate a good model fit, IRT-grm parameter and standard error estimates should examined to assess the scaling properties.

For IRT-mok the scalability coefficient $H$ on scale level is commonly used to assess model fit. Scale $H$ was found to be a useful indicator of scale strength regardless of the LV or item distributions. Only for scales consisting of items with equal means, $H$ suffered from a small positive bias, as did the item $H_i$ coefficients.

It should be noted that the behavior of fit statistics in case of model *mis*fit was not investigated in our study.

#### Use Model-Estimated LV Scores

Regardless of the model applied, many researchers use unweighted sum scores to represent the scale scores and thus the respondents' LV scores. This practice is not recommended because, consequently, valuable information about the LV distribution is lost. Furthermore, inferences about the shape of the LV distribution can be faulty when based on unweighted sum scores rather than model-estimated LV scores. When applying a parametric model, it is advised to use the information about parameter estimates by using the model-estimated LV scores.

## 8.3 Qualifications of the Monte Carlo Study

We compared the behavior of four estimation models, FA-lin, FA-poly, IRT-grm, and IRT-mok. The population model used for data generation in the Monte Carlo design was FA-poly or, equivalently, the IRT-grm model, thus resulting in ordered categorical data. From a theoretical point of view, it was therefore to be expected that FA-lin parameter

estimators would be biased. One could argue that the parameters estimated by FA-lin are essentially different from the true parameters they are compared with in our assessment of estimation performance. From a practitioner's point of view, however, it is important to know the extent to which FA-lin parameters are biased in case of ordered categorical data because, as was shown in Chapter 2, applying FA-lin to ordered categorical data is common practice.

In that regard, the ML estimation method was chosen for FA-lin, because it is the standard practice and the default method for FA-lin in software packages such as MPLUS. One should note that robust ML variants, such as MLR and mean-and-variance adjusted ML, only differ from ordinary ML with respect to the estimation of standard errors and model fit indices; parameter estimation is identical to ML.

The same reasoning holds for applying FA-poly and IRT-grm to data with a non-normal underlying LV. As both models assume a normal LV, one could expect some estimation problems when employing them in case that assumption is violated. The lack of normality of the LV is expected to cause some parameter bias in the threshold and loading parameters. From a practical perspective, it is relevant to assess the size of the estimation bias, or the robustness properties of each model against specific violations of assumptions. It was interesting to find that FA-poly is much more affected by a violation of the normality assumption than IRT-grm. This is probably due to the estimation method, with FA-poly taking into account only the univariate and bivariate properties of the data and IRT-grm using the full response patterns for parameter and standard error estimation.

The nonparametric IRT-mok does not make any other assumption than monotonically nondecreasing item-step response functions (ISRFs), i.e., with an increasing LV score the probability of endorsing an item must remain stable or increase. Therefore, nonnormal LV or item distributions were not expected to present any problems here. Although a comparison between parametric and nonparametric models is neither conventional nor straightforward, for example because direct comparisons of parameter estimation are not possible, we have included such comparisons in our study. Performance of IRT-mok was good, especially in case of nonnormal data. LV score estimates, however, were considered inferior to those of FA-poly and IRT-grm. It should be noted that we took unweighted sum scores to be the IRT-mok LV score estimates, as is commonly done in practice. Theoretically, however, IRT-mok merely provides an ordering of respondents. The shape of the LV distribution, which was discussed elaborately in our study, is theoretically undefined or unidentifiable in case of IRT-mok. Because in practice one might want to make inferences about the LV distribution, we compared the estimated LV distribution to its true counterpart, in addition to the comparison of the estimated and true ordering of respondents.

## 8.4 Suggestions for Future Research

It goes without saying that our Monte Carlo design did not include every possible empirical setting of importance. One is by default limited in the number of conditions

that can be thoroughly investigated. We chose a setup for an extensive examination of combinations of LV and item nonnormality, keeping many other design variables constant. Issues that were untouched in our study but are considered of special further interest are briefly discussed in the following.

### Nonnormal Mixed Scales

In the normal data configurations, items of various loadings were included, whereas item loadings were held constant at "strong" in the nonnormal data configurations. It is of interest to investigate mixed nonnormal scales, as FA-poly was found to be slightly superior in case of mixed loadings and a normal LV and IRT-grm outperformed FA-poly when the LV deviated from normality.

### Other Estimation Methods

In addition to the nonparametric IRT-mok, we compared three models, each with their own estimation method. FA-lin with ML estimation was included as the standard practice, although it is known to produce inferior results compared to other models available in case of ordered categorical data. When items are not specified to be categorical, it is the default method in MPLUS. FA-poly was presented as the more appropriate alternative for such data, with WLSMV estimation — which is equal to robust diagonally weighted least squares (cf. Yang-Wallentin et al., 2010) — considered to be the best method currently available. When items are identified as categorical, it is the default estimator in MPLUS. IRT-grm was included as the equivalent IRT formulation of FA-poly, with MLR estimation chosen as the best method currently available for ordered categorical data.

It should be noted that for each model other estimation methods could have been employed. We mention some methods that have shown good results for FA-poly: unweighted least squares (ULS) (e.g., Forero & Maydeu-Olivares, 2009; Forero et al., 2009), robust unweighted least squares and MLR (e.g., Yang-Wallentin et al., 2010), and Bayesian methods such as Markov chain Monte Carlo (e.g., Edwards, 2005). Bayesian modeling is also applied in IRT; see, e.g., J.-P. Fox (2010) for a reference textbook. It could be of interest to further investigate these and other estimation methods available for FA and IRT modeling.

### Multidimensionality

Our design was limited to unidimensional scales. As we found some notable differences between IRT-grm and the FA models for the multidimensional analysis briefly presented in Chapter 7, the comparison between FA-poly and IRT-grm is worth examining for the multidimensional case. IRT-grm loading parameter estimates differed between the unidimensional and the multidimensional case, whereas for both FA-lin and FA-poly differences were minimal. The only study (to our knowledge) including IRT-grm in a multidimensional design (i.e., Forero & Maydeu-Olivares, 2009) presented

results distinguishing between item/LV ratios rather than unidimensional versus multidimensional.

### Larger Skewness

Only one degree of nonnormal LV skewness was included in our study, as skewness was either 0 or 0.96. The skew-normal distribution, as proposed by Azzalini (1985), was employed with a shape parameter of 10 resulting in approximately the most skewed distribution possible for this family of distributions. Larger LV skewnesses, resulting from employing, e.g., a gamma distribution, were not investigated. With regard to nonnormal LV distributions, estimation models taking into account LV skewness (e.g., D. Molenaar, Dolan, & De Boeck, 2012) are also of interest to further investigations.

Larger skewness of item variables could be another topic of future research. The item skewness included in our study of 1.51 is moderate compared to some of the skewness values encountered in the empirical data used in the applications of Chapter 7, where values between 2.5 and 3.0 were not uncommon.

### Intermediate Sample Size

Since we found precision of parameter estimators to be low in case of the small sample size of $n = 200$ and satisfactory in case of $n = 600$, the question arises whether an intermediate sample size, of e.g., $n = 350$, would suffice for robust estimation performance.

### Model Misspecification

As for the structure of the models, no model misspecification was included in our design. The combination of a misspecified model and nonnormal data configurations is well worth investigating. One type of model misfit was investigated by Finch (2011), who applied a simple-structure model to data generated according to a semi-complex and complex structure. Finch found FA of the estimated tetrachoric correlation matrix (FA-tet) to produce lower parameter bias than two-parameter IRT model (IRT-2p) in such cases. Yang-Wallentin et al. (2010) examined the comparative performance of a number of FA-poly estimation methods in case of model misspecification. Skewed item distributions were included in the design, but the LV distribution was kept normal. It would be interesting to investigate the performance of FA-poly in case of a nonnormal LV under conditions of model misspecification, and compare it to IRT-grm.

### Missing Data

Missing data were not included in our simulation study. The applications of Chapter 7, however, clearly demonstrated the relevance of this issue. In each of these data sets observations were missing, and in each application missing data were handled differently, corresponding to the actions employed in the original analyses: adaptation of the estimation method to full-information maximum likelihood in MPLUS, listwise

deletion, and multiple imputation. The effect of various types of missing data and methods employed for handling missing data on model estimation results is certainly an interesting topic for future research.

### Practical Concluding Remarks

Many questions regarding the robustness of FA and IRT against violations of distributional assumptions have been answered in this dissertation. Using our general approach, the additional research questions suggested here can readily be investigated. The R scripts used to generate the data for the Monte Carlo study, as well as the R code for the implemented fit statistics ($\chi^2_{YB}$, $\chi^2_{YB}$-based RMSEA, and SRMR) are available online[a]. The setup of the MPLUS scripts is reported in Appendix C.4. All could be used as a convenient starting point to accommodate further research questions.

[a] See `http://irs.ub.rug.nl/data/1` at the University of Groningen library domain.

# References

References marked with an asterisk indicate studies included in the review of Chapter 2.


Adams, R. J. (2005). Reliability as a measurement design effect. *Studies in Educational Evaluation*, *31*, 162–172.

*Aluja, A., Blanch, A., & García, L. F. (2005). Dimensionality of the Maslach Burnout Inventory in school teachers: A study of several proposals. *European Journal of Psychological Assessment*, *21*, 67–76.

American Psychological Association. (2001). *Publication manual of the American Psychological Association* (5th ed.). Washington, DC: Author.

American Psychological Association. (2010). *Publication manual of the American Psychological Association* (6th ed.). Washington, DC: Author.

Andersen, P. A., Eloy, S. V., Guerrero, L. K., & Spitzberg, B. H. (1995). Romantic jealousy and relational satisfaction: A look at the impact of jealousy experience and expression. *Communication Reports*, *8*, 77–85.

Andrich, D. (1997). A hyperbolic cosine irt model for unfolding direct responses of persons to items. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 399–414). New York, NY: Springer.

Arbuckle, J. (1995–2006). Amos 7.0 user's guide [Computer software manual]. Chicago, IL: SPSS.

*Arrindell, W., Akkerman, A., Bagés, N., Feldman, L., Caballo, V. E., Oei, T. P., ... Zaldívar, F. (2005). The short-EMBU in Australia, Spain, and Venezuela. *European Journal of Psychological Assessment*, *21*, 56–66.

*Aycicegi, A., Dinn, W. M., & Harris, C. L. (2005). Validation of Turkish and English Versions of the Schizotypal Personality Questionnaire — B. *European Journal of Psychological Assessment*, *21*, 34–43.

Azzalini, A. (1985). A class of distributions which includes the normal ones. *Scandinavian Journal of Statistics*, *12*, 171–178.

Azzalini, A. (2005). The skew-normal distribution and related multivariate families. *Scandinavian Journal of Statistics*, *32*, 159–188.

Azzalini, A. (2007). R package `sn`: The skew-normal and skew-$t$ distributions (Version 0.4-4) [Computer software]. Retrieved from `http://azzalini.stat.unipd.it/SN`.

Babakus, E., Ferguson, C. E., & Jöreskog, K. G. (1987). The sensitivity of confirmatory maximum likelihood factor analysis to violations of measurement scale and distributional assumptions. *Journal of Marketing Research*, *24*, 222–228.

Bacon, D. R., Sauer, P. S., & Young, M. (1995). Composite reliability in structural equation modeling. *Educational and Psychological Measurement*, *55*, 394–406.

Baker, F. B., & Kim, S.-H. (2004). *Item response theory: Parameter estimation techniques* (2nd ed.). New York, NY: Dekker.

Bandalos, D. L., & Finney, S. J. (2010). Factor analysis: Exploratory and confirmatory. In G. R. Hancock & R. O. Mueller (Eds.), *The reviewer's guide to quantitative methods in the social sciences* (pp. 93–114). New York, NY: Routledge.

Barelds, D. P. H., & Barelds-Dijkstra, P. (2007). Relations between different types of jealousy and self and partner perceptions of relationship quality. *Clinical Psychology and Psychotherapy*, *14*, 176–188.

Barelds, D. P. H., & Dijkstra, P. (2003). Het meten van jaloezie [Measuring jealousy]. *Diagnostiekwijzer*, *6*, 56–67.

Barelds, D. P. H., & Dijkstra, P. (2006). Reactive, anxious, and possessive forms of jealousy and their relation to relationship quality among heterosexuals and homosexuals. *Journal of Homosexuality*, *51*, 183–198.

Beauducel, A., & Herzberg, P. Y. (2006). On the performance of maximum likelihood versus means and variance adjusted weighted least squares estimation in CFA. *Structural Equation Modeling*, *13*, 186–203.

Bejar, I. I. (1983). Introduction to item response models and their assumptions. In R. K. Hambleton (Ed.), *Applications of IRT* (pp. 1–23). Vancouver, Canada: Educational Research Institute of British Columbia.

Bentler, P. M. (1989). EQS structural equations program manual [Computer software manual]. Los Angeles, CA: BMDP Statistical Software.

Bentler, P. M. (1995). EQS program manual [Computer software manual]. Encino, CA: Multivariate Software.

Bentler, P. M. (2009). Alpha, dimension-free, and model-based internal consistency reliability. *Psychometrika*, *74*, 137–143.

Bernaards, C. A., & Sijtsma, K. (2000). Influences on imputation and EM methods on factor analysis when item nonresponse in questionnaire data is nonignorable. *Multivariate Behavioral Research*, *35*, 321–364.

*Beyers, W., Goossens, L., Calster, B. V., & Duriez, B. (2005). An alternative substantive factor structure of the Emotional Autonomy Scale. *European Journal of Psychological Assessment*, *21*, 147–155.

Bock, R. D., & Aitken, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*, 443–459.

Bock, R. D., Gibbons, R., & Muraki, E. (1988). Full-information item factor analysis. *Applied Psychological Measurement*, *12*, 261–280.

Bock, R. D., & Lieberman, M. (1970). Fitting a response model forn dichotomously scored items. *Psychometrika*, *35*, 179–197.

Boker, S., Neale, M., Maes, H., Wilde, M., Spiegel, M., Brick, T., ... Fox, J. (2011). OpenMx: An open source extended structural equation modeling framework. *Psychometrika*, *76*, 306–317.

Bollen, K. A. (1989). *Structural equations with latent variables*. New York, NY: Wiley.

Bollen, K. A. (2002). Latent variables in psychology and the social sciences. *Annual Review of Psychology*, *53*, 605–634.

Bollen, K. A., & Lennox, R. (1991). Conventional wisdom on measurement: A structural equation perspective. *Psychological Bulletin*, *110*, 305–314.

Bolt, D. M. (2005). Limited- and full-information estimation of item response theory models. In A. Maydeu-Olivares & J. J. McArdle (Eds.), *Contemporary psychometrics* (pp. 27–71). Mahwah, NJ: Erlbaum.

Boomsma, A. (1983). *On the robustness of LISREL (maximum likelihood estimation) against small sample size and non-normality*. Amsterdam: Sociometric Research Foundation (doctoral dissertation, University of Groningen).

Boomsma, A. (1985). Nonconvergence, improper solutions, and starting values in LISREL maximum likelihood estimation. *Psychometrika*, *50*, 229–242.

Boomsma, A. (2000). Reporting analyses of covariance structures. *Structural Equation Modeling*, *7*, 461–483.

Boulet, J. R. (1996). *The effect of nonnormal ability distributions on IRT parameter estimation using full-information and limited-information methods* (Unpublished doctoral dissertation). University of Ottawa, Canada.

Brown, M. B., & Benedetti, J. K. (1977). On the mean and variance of the tetrachoric correlation coefficient. *Psychometrika*, *42*, 347–355.

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York, NY: The Guilford Press.

Browne, M. W. (1984). Asymptotically distribution-free methods for the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology*, *37*, 62–83.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136–162). Beverly Hills, CA: Sage.

Buunk, B. P. (1982). Anticipated sexual jealousy: Its relationship to self-esteem, dependency, and reciprocity. *Personality and Social Psychology Bulletin*, *8*, 310–316.

Buunk, B. P. (1997). Personality, birth order and attachment styles as related to various types of jealousy. *Personality and Individual Differences*, *23*, 997–1006.

*Calvete, E., Estévez, A., de Arroyabe, E. L., & Ruiz, P. (2005). The Schema Questionnaire — Short Form. *European Journal of Psychological Assessment*, *21*, 90–99.

Camilli, G. (1994). Origin of the scaling constant $d = 1.7$ in item response theory. *Journal of Educational and Behavioral Statistics*, *19*, 293–295.

Camstra, A., & Boomsma, A. (1992). Cross-validation in regression and covariance structure analysis. *Sociological Methods & Research*, *21*, 89–115.

*Caprara, G. V., Steca, P., Zelli, A., & Capanna, C. (2005). A new scale for measuring adults' prosocialness. *European Journal of Psychological Assessment*, *21*, 77–89.

Carlson, J. E. (1987). Multidimensional item response theory estimation: A computer program [Computer software manual]. Retrieved from `http://www.act.org/research/reports/pdf/ACT_RR87-19.pdf`.

*Cashin, S. E., & Elmore, P. B. (2005). The Survey of Attitudes Toward Statistics Scale: A construct validity study. *Educational and Psychological Measurement*, *65*, 509–524.

Chen, F., Curran, P. J., Bollen, K. A., Kirby, J., & Paxton, P. (2008). An empirical evaluation of the use of fixed cutoff points in RMSEA test statistic in structural equation models. *Sociological Methods & Research*, *36*, 462–494.

Chen, W.-H. (1993). IRT-LD: A computer program for the detection of pairwise local dependence between test items (Research memorandum 93-2) [Computer software manual]. Chapel Hill: L. L. Thurstone Laboratory, University of North Carolina at Chapel Hill.

Christoffersson, A. (1975). Factor analysis of dichotomized variables. *Psychometrika*, *40*, 5–32.

*Clark, D. A., Antony, M. M., Beck, A. T., Swinson, R. P., & Steer, R. A. (2005). Screening for obsessive and compulsive symptoms: Validation of the Clark-Beck Obsessive-Compulsive Inventory. *Psychological Assessment*, *17*, 132–143.

Coenders, G., Satorra, A., & Saris, W. E. (1997). Alternative approaches to structural modeling of ordinal data: A Monte Carlo study. *Structural Equation Modeling*, *4*, 261–282.

Cohen, J. (1973). Eta-squares and partial eta-squared in fixed factor anova designs. *Educational and Psychological Measurement*, *33*, 107–112.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology*, *78*, 98–104.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*, 297–334.

Cudeck, R. (1989). Analysis of correlation matrices using covariance structure models. *Psychological Bulletin*, *105*, 317–327.

Cudeck, R., & Browne, M. W. (1983). Cross-validation of covariance structures. *Multivariate Behavioral Research*, *18*, 147–167.

De Ayala, R. J. (2010). Item response theory. In G. R. Hancock & R. O. Mueller (Eds.), *The reviewer's guide to quantitative methods in the social sciences* (pp. 155–171). New York, NY: Routledge.

De Champlain, A. F., & Tang, K. L. (1997). CHIDIM: A FORTRAN program for assessing the dimensionality of binary item responses based on McDonald's

nonlinear factor analytic model. *Educational and Psychological Measurement*, *33*, 181–201.

*De Frias, C. M., & Dixon, R. A. (2005). Confirmatory factor structure and measurement invariance of the Memory Compensation Questionnaire. *Psychological Assessment*, *17*, 168–178.

De Gruijter, D. N. M. (1994). Comparison of the nonparametric Mokken model and parametric IRT models using latent class analysis. *Applied Psychological Measurement*, *18*, 27–34.

DeMars, C. E. (2010, May). *A comparison of limited-information and full-information methods in Mplus for estimating IRT parameters for non-normal populations.* Paper presented at the annual meeting of the National Council on Measurement in Education, Denver.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B*, *39*, 1–38.

DiStefano, C. (2002). The impact of categorization with confirmatory factor analysis. *Structural Equation Modeling*, *9*, 327–346.

Dolan, C. V. (1994). Factor analysis of variables with 2, 3, 5 and 7 response categories: A comparison of categorical variable estimators using simulated data. *British Journal of Mathematical and Statistical Psychology*, *47*, 309–326.

Drasgow, F. (1989). An evaluation of marginal maximum likelihood estimation for the two-parameter logistic model. *Applied Psychological Measurement*, *13*, 77–90.

Dumenci, L., & Achenbach, T. M. (2008). Effects of estimation methods on making trait-level inferences from ordered categorical items for assessing psychopathology. *Psychological Assessment*, *20*, 55–62.

Eddelbuettel, D. (2009). `random` (Version 0.2.1): True random numbers using `random.org` [Computer software]. Retrieved from `http://CRAN.R-project.org/package=random`.

Edwards, M. C. (2005). *A markov chain monte carlo approach to confirmatory item factor analysis* (Unpublished doctoral dissertation). University of North Carolina at Chapel Hill.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists.* Mahwah, NJ: Erlbaum.

Ferrando, P. J., & Lorenzo-Seva, U. (2000). Unrestricted versus restricted factor analysis of multidimensional test items: Some aspects of the problem and some suggestions. *Psicológica*, *21*, 301–323.

Finch, H. (2010). Item parameter estimation for the MIRT model: Bias and precision of confirmatory factor analysis-based models. *Applied Psychological Measurement*, *34*, 10–26.

Finch, H. (2011). Multidimensional item response theory parameter estimation with nonsimple structure items. *Applied Psychological Measurement*, *35*, 67–82.

Finger, M. S. (2001). *A comparison of full-information and unweighted least-squares limited-information methods used with the 2-parameter normal ogive model* (Unpublished doctoral dissertation). University of Minnesota.

*Fletcher, R., & Hattie, J. (2005). Gender differences in physical self-concept: A multidimensional differential item functioning analysis. *Educational and Psychological Measurement*, *65*, 657–667.

Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods*, *9*, 466–491.

Floyd, F. J., & Widaman, K. F. (1995). Factor analysis in the development and refinement of clinical assessment instruments. *Psychological Assessment*, *7*, 286–299.

Forero, C. G., & Maydeu-Olivares, A. (2009). Estimation of IRT graded response models: Limited versus full information methods. *Psychological Methods*, *14*, 275–299.

Forero, C. G., Maydeu-Olivares, A., & Gallordo-Pujol, D. (2009). Factor analysis with ordinal indicators: A Monte Carlo study comparing DWLS and ULS estimation. *Structural Equation Modeling*, *16*, 625–641.

Fox, J. (2006). Structural equation modeling with the sem package in R. *Structural Equation Modeling*, *13*, 465–486.

Fox, J., Nie, Z., & Byrnes, J. (2013). `sem` (R package Version 3.1-3): Structural equation models [Computer software]. Retrieved from `CRAN.R-project.org/package=sem`.

Fox, J.-P. (2010). *Bayesian item response modeling: Theory and applications.* New York, NY: Springer.

Fraser, C. (1983). NOHARM: A FORTRAN program for fitting unidimensional and multidimensional normal ogive models of latent trait theory [Computer software]. Armidale, Australia: University of England, Centre for Behavioural Studies.

Frieling, M. A. (2008). *Een goede buur: 'joint production' als motor voor actief burgerschap in de buurt* [Love thy neighbor... Increasing the communal strength of neighborhoods] (Unpublished doctoral dissertation). University of Groningen.

Gerbing, D. W., & Hamilton, J. G. (1996). Viability of exploratory factor analysis as a precursor to confirmatory factor analysis. *Structural Equation Modeling*, *3*, 62–72.

*Ghaderi, A. (2005). Psychometric properties of the Self-Concept Questionnaire. *European Journal of Psychological Assessment*, *21*, 139–146.

*Glutting, J. J., Watkins, M. W., & Youngstrom, E. A. (2005). ADHD and college students: Exploratory and confirmatory factor structures with student and parent data. *Psychological Assessment*, *17*, 44–55.

Green, B. F., Bock, D., Humphreys, L. G., Linn, R. L., & Reckase, M. D. (1984). Technical guidelines for assessing computerized adaptive tests. *Journal of Educational Measurement*, *21*, 347–360.

Green, S. B., & Yang, Y. (2009a). Commentary on coefficient alpha: A cautionary tale. *Psychometrika*, *74*, 121–135.

Green, S. B., & Yang, Y. (2009b). Reliability of summed item scores using structural equation modeling: An alternative to coefficient alpha. *Psychometrika*, *74*, 155–167.

*Grothe, K. B., Dutton, G. R., Jones, G. N., Bodenlos, J., Ancona, M., & Brantley, P. J. (2005). Validation of the Beck Depression Inventory — II in a low-income African American sample of medical outpatients. *Psychological Assessment*, *17*, 110–114.

Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika*, *23*, 297–308.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory*. Boston: Kluwer-Nijhof.

Hancock, G. R., & Mueller, R. O. (Eds.). (2010). *The reviewer's guide to quantitative methods in the social sciences*. New York, NY: Routledge.

Harman, H. H. (1968). *Modern factor analysis* (2nd ed.). Chicago, IL: University of Chicago.

Harman, H. H., & Jones, W. H. (1966). Factor analysis by minimizing residuals (MINRES). *Psychometrika*, *31*, 351–368.

Hatcher, L. (1994). *A step-by-step approach to using the SAS system for factor analysis and structural equation modeling*. Cary, NC: SAS Institute.

*Heinitz, K., Liepmann, D., & Felfe, J. (2005). Examining the factor structure of the MLQ: Recommendation for a reduced set of factors. *European Journal of Psychological Assessment*, *21*, 182–190.

Hemker, B. T., & Sijtsma, K. (1995). Selection of unidimensional scales from a multidimensional item bank in the polytomous Mokken IRT model. *Applied Psychological Measurement*, *19*, 337–352.

Henze, N. (1986). A probabilistic representation of the 'skew-normal' distribution. *Scandinavian Journal of Statistics*, *13*, 271–275.

Holgado-Tello, F. P., Chacón-Moscoso, S., Barbero-García, I., & Vila-Abad, E. (2010). Polychoric versus Pearson correlations in exploratory and confirmatory factor analysis of ordinal variables. *Quality & Quantity*, *44*, 153–166.

Holst, K. K., & Budtz-Joergensen, E. (2012). Linear latent variable models: The lava-package. *Computational Statistics*, *28*, 1385–1452.

Holzinger, K. J. (1944). A simple method of factor analysis. *Psychometrika*, *9*, 257–262.

*Hong, S., & Wong, E. C. (2005). Rasch rating scale modeling of the Korean version of the Beck Depression Inventory. *Educational and Psychological Measurement*, *65*, 124–139.

Hoogland, J. J. (1999). *The robustness of estimation methods for covariance structure analysis* (Unpublished doctoral dissertation). University of Groningen.

Hoskens, M., & De Boeck, P. (1997). A parametric model for local item dependencies among test items. *Psychological Methods*, *32*, 234–246.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, *24*, 417–441 and 498–520.

Hoyle, R. H., & Panter, A. T. (1995). Writing about structural equation models. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues and applications* (pp. 158–198). Thousand Oaks, CA: Sage.

Hu, L., & Bentler, P. M. (1995). Evaluating model fit. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues and applications* (pp. 76–99). Thousand Oaks, CA: Sage.

Hu, L., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods*, *4*, 424–453.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*, 1–55.

Hulin, C. L., Drasgow, F., & Parsons, C. K. (1983). *Item response theory.* Homewood, IL: Dorsey.

Iacobucci, D., & Duhachek, A. (2003). Advancing alpha: Measuring reliability with confidence. *Journal of Consumer Psychology*, *13*, 478–487.

*Inglés, C. J., Hidalgo, M. D., & Méndez, F. X. (2005). Interpersonal difficulties in adolescence: A new self-report measure. *European Journal of Psychological Assessment*, *21*, 11–22.

Jackson, P. H., & Agunwamba, C. C. (1977). Lower bounds for the reliability of the total score on a test composed of non-homogeneous items: I: Algebraic lower bounds. *Psychometrika*, *42*, 567–578.

Jöreskog, K. G. (1967). Some contributions to maximum likelihood factor analysis. *Psychometrika*, *32*, 443–482.

Jöreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika*, *34*, 183–202.

Jöreskog, K. G., & Goldberger, A. S. (1972). Factor analysis by generalized least squares. *Psychometrika*, *37*, 243–260.

Jöreskog, K. G., & Moustaki, I. (2001). Factor analysis of ordinal variables: A comparison of three approaches. *Multivariate Behavioral Research*, *36*, 347–387.

Jöreskog, K. G., & Sörbom, D. (1981). LISREL V: Analysis of linear structural relationships by the method of maximum likelihood [Computer software manual]. Chicago, IL: National Educational Resources.

Jöreskog, K. G., & Sörbom, D. (1986). LISREL VI: Analysis of linear structural relationships by maximum likelihood and least square methods [Computer software manual]. Mooresville, IN: Scientific Software International.

Jöreskog, K. G., & Sörbom, D. (1996). LISREL 8: User's reference guide [Computer software manual]. Chicago, IL: Scientific Software International.

*Joseph, S., Linley, P. A., Andrews, L., Harris, G., Howle, B., & Woodward, C. (2005). Assessing positive and negative changes in the aftermath of adversity:

Psychometric evaluation of the Changes in Outlook Questionnaire. *Psychological Assessment*, *17*, 70–80.

Kaiser, H. F., & Caffrey, J. (1965). Alpha factor analysis. *Psychometrika*, *30*, 1–14.

Kamata, A., & Bauer, D. J. (2008). A note on the relation between factor analytic and item response theory models. *Structural Equation Modeling*, *15*, 136–153.

Kay, C. A. (2004). *A comparison of traditional and IRT factor analysis* (Unpublished doctoral dissertation). University of North Texas.

Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika*, *30*, 81–93.

Kline, R. B. (1998). *Principles and practice of structural equation modeling.* New York, NY: The Guilford Press.

Kline, R. B. (2004). *Supplemental chapter: Multivariate effect size estimation.* Retrieved 2012/11/28, from {\url{http://forms.apa.org/books/supp/kline/pdfs/multivariate.pdf}}

Knol, D. L., & Berger, M. P. F. (1991). Empirical comparison between factor analysis and multidimensional item response models. *Multivariate Behavioral Research*, *26*, 457–477.

Koning, A. J., & Franses, P. H. (2003, June). *Confidence intervals for Cronbach's coefficient alpha values* (Tech. Rep. No. ERS-2003-041-MKT). Rotterdam, the Netherlands: Erasmus Research Institute of Management. Retrieved from http://publishing.eur.nl/ir/repub/asset/431/ERS-2003-041-MKT.pdf.

*Kotov, R., Schmidt, N. B., Zvolensky, M. J., Vinogradov, A., & Antipova, A. V. (2005). Adaptation of panic-related psychopathology measures to Russian. *Psychological Assessment*, *17*, 242–246.

Kuijpers, R. E., Van der Ark, L. A., & Croon, M. A. (2013). Standard errors and confidence intervals for scalability coefficients in Mokken scale analysis using marginal models. *Sociological Methodology*, *43*, 42–69.

*Le, H., Casillas, A., Robbins, S. B., & Langley, R. (2005). Motivational and skills, social, and self-management predictors of college outcomes: Constructing the Student Readiness Inventory. *Educational and Psychological Measurement*, *65*, 482–508.

*Leite, W. L., & Beretvas, S. N. (2005). Validation of scores on the Marlowe-Crowne Social Desirability Scale and the Balanced Inventory of Desirable Responding. *Educational and Psychological Measurement*, *65*, 140–154.

Levine, T. R., & Hullett, C. R. (2002). Eta squared, partial eta squared, and misreporting of effect size in communication research. *Human Communication Research*, *28*, 612–625.

Lindgren, B. W. (1993). *Statistical theory* (4th ed.). New York, NY: Chapman & Hall.

*Longley, S. L., Watson, D., & Noyes, R., Jr. (2005). Assessment of the hypochondriasis domain: The Multidimensional Inventory of Hypochondriacal Traits (MIHT). *Psychological Assessment*, *17*, 3–14.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

*Lowe, P. A., & Reynolds, C. R. (2005). Factor structure of AMAS-C scores across gender among students in collegiate settings. *Educational and Psychological Measurement*, *65*, 687–708.

*Marsh, H. W., Ellis, L. A., Parada, R. H., Richards, G., & Heubeck, B. (2005). A short version of the Self Description Questionnaire II: Operationalizing criteria for short-form evaluation with new applications of confirmatory factor analyses. *Psychological Assessment*, *17*, 81–102.

Marsh, H. W., & Hau, K.-T. (1999). Confirmatory factor analysis: Strategies for small sample sizes. In R. H. Hoyle (Ed.), *Statistical strategies for small sample research* (pp. 251–306). Thousand Oaks, CA: Sage.

Marsh, H. W., Hau, K.-T., Balla, J. R., & Grayson, D. (1998). Is more ever too much? The number of indicators per factor in confirmatory factor analysis. *Multivariate Behavioral Research*, *33*, 181–220.

Marsh, H. W., Hau, K.-T., & Wen, Z. (2004). In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings. *Structural Equation Modeling*, *11*, 320–341.

Martin, A. D., Quinn, K. M., & Park, J. H. (2007). MCMCpack: Markov chain Monte Carlo (MCMC) package. R package Version 0.9-2 [Computer software]. Retrieved from `http://mcmcpack.wustl.edu`.

Matsumoto, M., & Nishimura, T. (1998). Mersenne Twister: A 623-dimensionally equidistributed uniform pseudo-random number generator. *ACM Transactions on Modeling and Computer Simulations*, *8*, 3–30.

Maydeu-Olivares, A., Cai, L., & Hernández, A. (2011). Comparing the fit of item response theory and factor analysis models. *Structural Equation Modeling*, *18*, 333–356.

Maydeu-Olivares, A., Drasgow, F., & Mead, A. D. (1994). Distinguishing among parametric item response models for polychotomous ordered data. *Applied Psychological Measurement*, *18*, 245–256.

Maydeu-Olivares, A., & McArdle, J. J. (Eds.). (2005). *Contemporary psychometrics*. Mahwah, NJ: Erlbaum.

McCrae, R. R., Zonderman, A. B., Costa, P. T., Jr., Bond, M. H., & Paunonen, S. V. (1996). Evaluating replicability of factors in the Revised NEO Personality Inventory: Confirmatory factor analysis versus Procrustes rotation. *Journal of Personality and Social Psychology*, *70*, 552–566.

McDonald, R. P. (1967). Nonlinear factor analysis. *Psychometric Monograph No. 15*.

McDonald, R. P. (1970). The theoretical foundations of principal factor analysis, canonical factor analysis, and alpha factor analysis. *British Journal of Mathematical and Statistical Psychology*, *23*, 1–21.

McDonald, R. P. (1981). The dimensionality of tests and items. *British Journal of Mathematical and Statistical Psychology*, *34*, 100–117.

McDonald, R. P. (1997). Normal-ogive multidimensional model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 257–269). New York, NY: Springer.

McDonald, R. P., & Ho, M.-H. R. (2002). Principles and practice in reporting structural equation models. *Psychological Methods*, *7*, 64–82.

McKinley, R. L., & Reckase, M. D. (1983). MAXLOG: A computer program for the estimation of the parameters of a multidimensional logistic model [Computer software manual]. *Behavior Research Methods & Instrumentation*, *15*, 389–390.

Meade, A. W., & Lautenschlager, G. J. (2004). A comparison of item response theory and confirmatory factor analytic methodologies for establishing measurement equivalence/invariance. *Organizational Research Methods*, *7*, 361–388.

Mehta, P. D., & Taylor, W. P. (2006, June). *On the relationship between item response theory and factor analysis of ordinal variables: Multiple group case.* Paper presented at the 71st annual meeting of the Psychometric Society, HEC Montreal, Canada.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin*, *105*, 156–166.

Mislevy, R. J., & Bock, R. D. (1984). BILOG: Maximum likelihood item analysis and test scoring with logistic models [Computer software]. Mooresville, IN: Scientific Software.

Mokken, R. J. (1971). *A theory and procedure of scale analysis with applications in political research.* New York/Berlin: De Gruyter/Mouton.

Molenaar, D., Dolan, C. V., & De Boeck, P. (2012). The heteroscedastic graded response model with a skewed latent trait: Testing statistical and substantive hypotheses related to skewed item category functions. *Psychometrika*, *77*, 455–478.

Molenaar, I. W. (1974). De logistische en de normale kromme [The logistic and the normal curve]. *Nederlands Tijdschrift voor de Psychologie*, *29*, 415–420.

Molenaar, I. W. (1982). Mokken scaling revisited. *Kwantitatieve Methoden*, *8*, 145–164.

Molenaar, I. W. (1988). Formal statistics and informal data analysis, or why laziness should be discouraged. *Statistica Neerlandica*, *42*, 83–90.

Molenaar, I. W. (1991). A weighted Loevinger H-coefficient extending Mokken scaling to multicategory items. *Kwantitatieve Methoden*, *12*, 97–117.

Moustaki, I., Jöreskog, K. G., & Mavridis, D. (2004). Factor models for ordinal variables with covariate effects on the manifest and latent variables: A comparison of LISREL and IRT approaches. *Structural Equation Modeling*, *11*, 487–513.

Mroch, A. A., & Bolt, D. M. (2006). A simulation comparison of parametric and nonparametric dimensionality detection procedures. *Applied Measurement in Education*, *19*, 67–91.

Mueller, R. O. (1996). *Basic principles of structural equation modeling: An introduction to LISREL and EQS.* New York, NY: Springer.

Mulaik, S. A. (2010). *The foundations of factor analysis* (2nd ed.). Boca Raton, FL: Chapman & Hall/CRC Press.

*Muller, J. J., Creed, P. A., Waters, L. E., & Machin, M. A. (2005). The development and preliminary testing of a scale to measure the latent and manifest benefits of employment. *European Journal of Psychological Assessment*, *21*, 191–198.

Muthén, B. O. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika*, *49*, 115–132.

Muthén, B. O. (1987). LISCOMP: Analysis of linear structural equations with a comprehensive measurement model [Computer software manual]. Mooresville, IN: Scientific Software.

Muthén, B. O. (1998–2004). *Mplus technical appendices*. Los Angeles, CA: Muthén & Muthén.

Muthén, B. O. (2006). *IRT in Mplus*. Retrieved from `http://www.statmodel.com/download/MplusIRT2.pdf`.

Muthén, B. O., & Asparouhov, T. (2002). Latent variable analysis with categorical outcomes: Multiple-group and growth modeling in Mplus. *Mplus web notes*, *4*. Retrieved from `http://www.statmodel.com/download/webnotes/CatMGLong.pdf`.

Muthén, B. O., & Kaplan, D. (1985). A comparison of some methodologies for the factor analysis of non-normal Likert variables. *British Journal of Mathematical and Statistical Psychology*, *38*, 171–189.

Muthén, B. O., & Kaplan, D. (1992). A comparison of some methodologies for the factor analysis of non-normal Likert variables: A note on the size of the model. *British Journal of Mathematical and Statistical Psychology*, *45*, 19–30.

Muthén, L. K., & Muthén, B. O. (1998–2010). Mplus user's guide (6th ed.) [Computer software manual]. Los Angeles, CA: Author.

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York, NY: McGraw-Hill.

Olejnik, S., & Algina, J. (2000). Measures of effect size for comparative studies: Applications, interpretations, and limitations. *Contemporary Educational Psychology*, *25*, 241–286.

Olsson, U. (1979). Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika*, *44*, 443–460.

OpenMx Development Team. (2010). OpenMx documentation [Computer software manual]. Retrieved from `http://openmx.psyc.virginia.edu/docs/OpenMx/latest/OpenMxUserGuide.pdf`.

Osborne, J. W. (2008). Sweating the small stuff in educational psychology: how effect size and power reporting failed to change from 1969 to 1999, and what that means for the future of changing practices. *Educational Psychology*, *28*, 151–160. doi: doi: 10.1080/01443410701491718

Parry, C. D. H., & McArdle, J. J. (1991). An applied comparison of methods for least-squares factor analysis of dichotomous variables. *Applied Psychological Measurement*, *15*, 35–46.

Parshall, C. G., Kromrey, J. D., Chason, W. M., & Yi, Q. (1997, June). *Evaluation of parameter estimation under modified IRT models and small samples.* Paper presented at the annual meeting of the Psychometric Society, Gatlinburg, TN.

Pearson, K., & Pearson, E. S. (1922). On polychoric coefficients of correlation. *Biometrika*, *14*, 127–156.

*Pett, M. A., & Johnson, M. J. M. (2005). Development and psychometric evaluation of the Revised University Student Hassles Scale. *Educational and Psychological Measurement*, *65*, 984–1010.

Pöhlmann, K., Thiel, P., & Joraschky, P. (2008). Entwicklung und validierung des Dresdner Körperbildfragebogens (DKB-35) [Development and validation of the Dresden Body Image Questionnaire (DBIQ-35)]. In P. Joraschky, H. Lausberg, & K. Pöhlmann (Eds.), *Körperorientierte Diagnostik und Psychotherapie bei Patientinnen mit Essstörungen* (pp. 57–72). Gießen, Germany: Psychosozial-Verlag.

Potthast, M. J. (1993). Confirmatory factor analysis of ordered categorical variables with large models. *British Journal of Mathematical and Statistical Psychology*, *46*, 273–286.

Puente, S., & Cohen, D. (2003). Jealousy and the meaning (or nonmeaning) of violence. *Psychological Bulletin*, *29*, 449–460.

R Core Team. (2013). R: A language and environment for statistical computing [Computer software]. Vienna: R Foundation for Statistical Computing. Retrieved from `http://www.R-project.org/`

Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2004). GLLAMM manual. U.C. Berkeley division of biostatistics working paper series. Working paper 160 [Computer software manual]. Retrieved from `http://www.bepress.com/ucbbiostat/paper160`.

Ramsay, J. O. (2000). TestGraf: A program for the graphical analysis of multiple choice test and questionnaire data [Computer software manual]. Retrieved from `ftp://ego.psych.mcgill.ca/pub/ramsay/testgraf/`.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: The Danish Institute of Educational Research.

Raykov, T. (1998). Coefficient alpha and composite reliability with interrelated nonhomogeneous items. *Applied Psychological Measurement*, *22*, 375–385.

Raykov, T., & Shrout, P. E. (2002). Reliability of scales with general structure: Point and interval estimation using a structural equation modeling approach. *Structural Equation Modeling*, *9*, 195–212.

Reckase, M. D. (2009). *Multidimensional item response theory.* New York, NY: Springer.

Revelle, W., & Zinbarg, R. E. (2009). Coefficients alpha, beta, omega, and the glb: Comments on Sijtsma. *Psychometrika*, *74*, 145–154.

Rhemtulla, M., Brosseau-Liard, P. É., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and cat-

egorical SEM estimation methods under suboptimal conditions. *Psychological Methods*, *17*, 354–373.

*Rivas, T., Bersabé, R., & Berrocal, C. (2005). Application of the Double Monotonicity Model to polytomous items, scalability of the Beck depression items on subjects with eating disorders. *European Journal of Psychological Assessment*, *21*, 1–10.

Rosseel, Y. (2012). `lavaan`: An R package for structural equation modeling. *Journal of Statistical Software*, *48*, 1–36.

Roussos, L. (1995). Hierarchical agglomerative clustering computer program user's manual [Computer software manual]. Urbana-Champaign: Statistical Laboratory for Educational and Psychological Measurement, Department of Statistics, University of Illinois.

*Sabourin, S., Valois, P., & Lussier, Y. (2005). Development and validation of a brief version of the Dyadic Adjustment Scale with a nonparametric item analysis model. *Psychological Assessment*, *17*, 15–27.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph*, *No. 17*.

Saris, W. E. (2008, September). *Tests of structural equation models do not work: What to do?* Paper presented at the 7th International Conference on Social Science Methodology, Naples, Italy.

Satorra, A., & Saris, W. E. (1985). Power of the likelihood ratio test in covariance structure analysis. *Psychometrika*, *50*, 83–90.

Scargle, J. D. (2000). Publication bias: The "file-drawer" problem in scientific inference. *Journal of Scientific Exploration*, *14*, 91–106.

Scheffers, M., Van Duijn, M. A. J., Bosscher, R. J., Wiersma, D., & Van Busschbach, J. T. (2013). *Psychometric properties of the Dutch translation of the Dresden Body Image Questionnaire: A multiple-group confirmatory factor analysis across sex and age in a non-clinical sample.* Manuscript submitted for publication.

Shackelford, T. K., & Buss, D. M. (2000). Marital satisfaction and spousal cost-infliction. *Personality and Individual Differences*, *28*, 917–928.

*Shevlin, M., & Adamson, G. (2005). Alternative factor models and factorial invariance of the GHQ-12: A large sample analysis using confirmatory factor analysis. *Psychological Assessment*, *17*, 231–236.

Sijtsma, K. (2009a). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika*, *74*, 107–120.

Sijtsma, K. (2009b). Reliability beyond theory and into practice. *Psychometrika*, *74*, 169–173.

Sijtsma, K., & Molenaar, I. W. (2002). *Introduction to nonparametric item response theory.* Thousand Oaks, CA: Sage.

*Simms, L. J., Casillas, A., Clark, L. A., Watson, D., & Doebbeling, B. N. (2005). Psychometric evaluation of the restructured clinical scales of the MMPI — 2. *Psychological Assessment*, *17*, 345–358.

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models.* New York, NY: Chapman & Hall.

Song, W. T., & Schmeiser, B. W. (2008). Displaying statistical point estimators: The leading-digit procedure. In S. J. Mason, R. R. Hill, L. Mönch, O. Rose, T. Jefferson, & J. W. Fowler (Eds.), *Proceedings of the 2008 winter simulation conference* (pp. 407–412). Miami, FL: WSC.

Song, W. T., & Schmeiser, B. W. (2009). Omitting meaningless digits in point estimates: The probability guarantee of leading-digit rules. *Operations Research*, *57*, 109–117.

Spearman, C. (1904). General intelligence, objectively determined and measured. *American Journal of Psychology*, *15*, 201–293.

Steiger, J. H. (1990). Structural model evaluation and modification: An interval estimation approach. *Multivariate Behavioral Research*, *25*, 173–180.

Steiger, J. H., & Lind, J. C. (1980, May). *Statistically based tests for the number of common factors.* Paper presented at the annual Spring Meeting of the Psychometric Society, Iowa City, IA.

*Stepleman, L. M., Darcy, M. U., & Tracey, T. J. (2005). Helping and coping attributions: Development of the Attribution of Problem Cause and Solution Scale. *Educational and Psychological Measurement*, *65*, 525–542.

Stevens, J. P. (2002). *Applied multivariate statistics for the social sciences* (4th ed.). Mahwah, NJ: Erlbaum.

Stone, C. A. (1992). Recovery of marginal maximum likelihood estimates in the two-parameter logistic response models: An evaluation of MULTILOG. *Applied Psychological Measurement*, *16*, 1–16.

Stout, W., Douglas, J., Junker, B., & Roussos, L. (1993). DIMTEST manual [Computer software manual]. Urbana-Champaign: University of Illinois at Urbana-Champaign, Department of Statistics.

Stuart, A., & Ord, J. K. (1987). *Kendall's advanced theory of statistics: Vol. 1. Distribution theory* (5th ed.). London: Griffin.

Tabachnick, B. G., & Fidell, L. S. (1983). *Using multivariate statistics.* New York, NY: Harper & Row.

Takane, Y., & De Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika*, *52*, 393–408.

Tate, R. (2003). A comparison of selected empirical methods for assessing the structure of responses to test items. *Applied Psychological Measurement*, *27*, 159–203.

Thurstone, L. L. (1947). *Multiple-factor analysis.* Chicago, IL: University of Chicago Press.

*Toland, M. D., & De Ayala, R. J. (2005). A multilevel factor analysis of students' evaluations of teaching. *Educational and Psychological Measurement*, *65*, 272–296.

*Tomás-Sábado, J., & Gómez-Benito, J. (2005). Construction and validation of the Death Anxiety Inventory (DAI). *European Journal of Psychological Assessment*, *21*, 108–114.

*Torff, B., Sessions, D., & Byrnes, K. (2005). Assessment of teachers' attitudes about professional development. *Educational and Psychological Measurement*, *65*, 820–830.

Trierweiler, T. (2009). *An evaluation of estimation methods in confirmatory factor analytic models with ordered categorical data in LISREL* (Unpublished doctoral dissertation). Fordham University, New York, NY.

Tuerlinckx, F., & De Boeck, P. (2001). The effect of ignoring item interactions on the estimated discrimination parameters in item response theory. *Psychological Methods*, *6*, 181–195.

Vacha-Haase, T. (1998). Reliability generalization: Exploring variance in measurement error affecting score reliability across studies. *Educational and Psychological Measurement*, *58*, 6–20.

Van Abswoude, A. A. H., Van der Ark, L. A., & Sijtsma, K. (2004). A comparative study of test data dimensionality assessment procedures under nonparametric IRT models. *Applied Psychological Measurement*, *28*, 3–24.

Van der Ark, L. A. (2001). Relationships and properties of polytomous item response theory models. *Applied Psychological Measurement*, *25*, 273–282.

Van der Ark, L. A. (2007). Mokken scale analysis in R. *Journal of Statistical Software*, *20*, 1–19.

Van der Ark, L. A. (2011). `mokken` (Version 2.7): An R package for Mokken scale analysis [Computer software]. Retrieved from `http://cran.r-project.org/web/packages/mokken/index.html`.

Van der Linden, W. J., & Hambleton, R. K. (Eds.). (1997). *Handbook of modern item response theory*. New York, NY: Springer.

*Van der Pas, S., Van Tilburg, T., & Knipscheer, K. C. P. M. (2005). Measuring older adults' filial responsibility expectations: Exploring the application of a vignette technique and an item scale. *Educational and Psychological Measurement*, *65*, 1026–1045.

Van Duijn, M. A. J. (1993). *Mixed models for repeated count data*. Leiden, the Netherlands: DSWO.

Van Onna, M. J. H. (2004). Estimates of the sampling distribution of scalability coefficient H. *Applied Psychological Measurement*, *28*, 427–449.

*Verdugo, M. A., Prieto, G., Caballo, C., & Peláez, A. (2005). Factorial structure of the Quality of Life Questionnaire in a Spanish sample of visually disabled adults. *European Journal of Psychological Assessment*, *21*, 44–55.

Vernon, T., & Eysenck, S. (Eds.). (2007). Special issue on structural equation modeling. *Personality and Individual Differences*, *42*(5).

Verschuren, P. J. M. (1991). *Structurele modellen tussen theorie en praktijk* [Structural models between theory and practice]. Utrecht, the Netherlands: Het Spectrum.

*Vigneau, F., & Bors, D. A. (2005). Items in context: Assessing the dimensionality of Raven's Advanced Progressive Matrices. *Educational and Psychological Measurement*, *65*, 109–123.

*Wang, M., & Russell, S. S. (2005). Measurement equivalence of the Job Descriptive Index across Chinese and American workers: Results from confirmatory factor analysis and item response theory. *Educational and Psychological Measurement*, *65*, 709–732.

*Weeks, J. W., Heimberg, R. G., Fresco, D. M., Hart, T. A., Turk, C. L., Schneier, F. R., & Liebowitz, M. R. (2005). Empirical validation and psychometric evaluation of the brief Fear of Negative Evaluation Scale in patients with social anxiety disorder. *Psychological Assessment*, *17*, 179–190.

*Wiebe, J. S., & Penley, J. A. (2005). A psychometric comparison of the Beck Depression Inventory–II in English and Spanish. *Psychological Assessment*, *17*, 481–485.

Wilkinson, L., & the Task Force on Statistical Inference. (1999). Statistical methods in psychology journals. *American Psychologist*, *54*, 594–604.

Wilson, D., Wood, R., & Gibbons, R. D. (1984). TESTFACT: Test scoring, item statistics, and item factor analysis [Computer software]. Mooresville, IN: Scientific Software.

Wu, M. L., Adams, R. J., & Wilson, M. R. (1998). ConQuest: Multi-aspect test software [Computer software manual]. Melbourne, Australia: Australian Council for Educational Research.

Yang-Wallentin, F., Jöreskog, K. G., & Luo, H. (2010). Confirmatory factor analysis of ordinal variables with misspecified models. *Structural Equation Modeling*, *17*, 392–423.

Yuan, K.-H., & Bentler, P. M. (1997). Mean and covariance structure analysis: Theoretical and practical improvements. *Journal of the American Statistical Association*, *92*, 767–774.

Yuan, K.-H., & Bentler, P. M. (1999). $F$ tests for mean and covariance structure analysis. *Journal of Educational and Behavioral Statistics*, *24*, 225–243.

Yuan, K.-H., & Bentler, P. M. (2000). Three likelihood-based methods for mean and covariance structure analysis with nonnormal missing data. *Sociological Methodology*, *30*, 165–200.

*Zapf, P. A., Skeem, J. L., & Golding, S. L. (2005). Factor structure and validity of the MacArthur Competence Assessment Tool — Criminal Adjudication. *Psychological Assessment*, *17*, 433–445.

Zegers, F. E., & Ten Berge, J. M. F. (1983). A fast and simple computational method of minimum residual factor analysis. *Multivariate Behavioral Research*, *18*, 331–340.

Zhang, J., & Stout, W. (1999). The theoretical DETECT index of dimensionality and its application to approximate simple structure. *Psychometrika*, *64*, 213–249.

*Zimprich, D., Perren, S., & Hornung, R. (2005). A two-level confirmatory factor analysis of a modified Rosenberg Self-Esteem Scale. *Educational and Psychological Measurement*, *65*, 465–481.

Zinbarg, R. E., Revelle, W., Yovel, I., & Li, W. (2005). Cronbach's $\alpha$, Revelle's $\beta$, and McDonald's $\omega_H$: Their relations with each other and two alternative conceptualizations of reliability. *Psychometrika*, *70*, 123–133.

# Appendix A

# Abbreviations

The numbers behind each abbreviation indicate the pages on which it is used.

**AGFI** adjusted goodness-of-fit index 37, 68

**AIC** Akaike's information criterion 38, 215

**ANOVA** analysis of variance vii, 48f., 52, 56, 68, 90f., 103f., 129f., 136, 144, 146, 167f., 180f.

**CA** component analysis 7f.

**CFA** confirmatory FA 30, 32ff., 37f., 40, 42

**CFI** comparative fit index 37, 63, 68, 216ff., 235f., 245, 367

**CTT** classical test theory 28, 30

**CVM** a categorical variable methodology 61, 70

**DBIQ** Dresden Body Image Questionnaire viif., 197ff., 213ff., 219ff., 238, 244, 247, 363ff.

**DIF** differential item functioning 29ff., 55

**DWLS** diagonally weighted least squares 11, 61, 70

**EAP** expected a posteriori 19, 39

**EFA** exploratory FA 30, 32ff., 37f., 40, 42

**EFA-tet-ULS** exploratory FA-tet by means of ULS 50, 68

**EJPA** *European Journal of Psychological Assessment* 28, 41

**EM** expectation maximization 18f., 46

**EPM** *Educational and Psychological Measurement* 28, 41

**FA** factor analysis vff., 1ff., 5, 7ff., 12, 16f., 20ff., 27ff., 38, 40ff., 49ff., 53ff., 63, 69ff., 75f., 80f., 84f., 91, 123, 130, 132, 141, 161, 173, 190, 197f., 200, 202, 204, 206, 208, 210, 212, 214, 216, 218, 220, 222ff., 226, 228, 230, 232, 234, 236, 238, 240, 242, 244, 246, 248, 250ff., 254, 258, 260, 296, 363ff.

**FA-lin** FA of the sample covariance matrix 2, 9ff., 24, 60, 69f., 72f., 75, 84f., 88f., 91ff., 98, 101, 103ff., 112ff., 123, 128ff., 132, 137ff., 143, 146, 148ff., 161, 165ff., 173, 180f., 184ff., 197f., 206, 212ff., 222, 229, 234ff., 238, 242, 244f., 247ff., 252ff., 294, 300ff., 306, 311ff., 315, 317, 319, 321, 323, 325, 327, 329ff., 333, 335, 337, 339, 341, 343, 345, 348ff., 360f., 367

**FA-pa** FA by means of a polynomial approximation of the normal-ogive function 59

**FA-poly** FA of the estimated polychoric correlation matrix viii, 2, 10ff., 21, 24f., 59ff., 69f., 72f., 75, 79, 84f., 88f., 91ff., 98, 101, 103, 105ff., 112ff., 132, 137ff., 143, 146, 149ff., 158ff., 173, 180f., 184ff., 195, 197f., 206, 212ff., 222f., 229, 234ff., 238, 242, 244f., 247ff., 252ff., 291, 294f., 300ff., 304, 306, 311f., 314, 316, 318, 320, 322, 324, 326, 328ff., 332, 334, 336, 338, 340, 342, 344, 346, 348ff., 357, 360f., 363ff.

**FA-tet** FA of the estimated tetrachoric correlation matrix 46, 59, 259

**FI** full-information 23ff., 251, 369

**FA-lin-ML** FA-lin by means of ML 2, 24, 57f., 60f., 68ff., 101, 142f., 252

**FA-pa-ULS** FA by means of a polynomial approximation of the normal-ogive function using ULS 46ff., 54ff., 59, 68

**FA-poly-CVM** FA-poly by means of a categorical variable methodology 62, 68

**FA-poly-DWLS** FA-poly by means of diagonally weighted least squares 62f., 68

**FA-poly-GLS** FA-poly by means of generalized least squares 61f., 68

**FA-phi-ULS** FA of the phi correlation matrix by means of ULS 49f., 68

**FA-poly-ML** FA-poly by means of ML 61f., 68

**FA-poly-MLR** FA-poly by means of MLR 63, 68

**FA-poly-ULS** FA-poly using ULS 52f., 61f., 68

**FA-poly-WLS** FA-poly by means of weighted least squares 62, 68

**FA-poly-WLSMV** FA-poly by means of WLSMV 2, 24, 52f., 58f., 61, 63, 68, 70f., 101, 142f., 252

**FA-tet-$\alpha$** alpha FA-tet 46, 68

**FA-tet-GLS** FA-tet by means of generalized least squares 46, 68

**FA-tet-IP** iterative principal FA-tet 46, 68

**FA-tet-MINR$_{adj}$** FA-tet by means of an adjusted minimum residuals method 46, 68

**FA-tet-ML** FA-tet by means of ML 46, 51f., 59, 68

**FA-tet-ULS** FA-tet by means of ULS 46, 49, 51, 62, 68

**FA-tet-WLSMV** FA-tet by means of WLSMV 50f., 54ff., 59, 68

**GFI** goodness-of-fit index 37, 61f., 68

**GLS** generalized least squares 23, 61

**IRT-2p-JML** IRT-2p by means of joint ML 46f., 64, 68

**IRT-2p-MLR** IRT-2p by means of MLR 54f., 59, 68

**IRT-2p-MML** IRT-2p by means of marginal ML 46ff., 56f., 59, 64, 66, 68

**IRT-3p-MML** the three-parameter IRT model by means of marginal ML 50f., 68

**IRT-grm-EAP** the graded response model by means of expected a posteriori estimation 58

**IRT-grm-ML** the graded response model by means of ML 52ff., 57f., 65, 68

**IRT-grm-MLR** the graded response model by means of MLR 2, 24, 71, 101, 142f., 252

**INCS** Involvement in Neighbourhood Community Scale ix, 197, 239ff., 247f., 368

**IRT-np-DET** nonparametric simple structure detection 50, 68

**IRT-np-DIM** nonparametric dimensionality testing 50, 68

**IRT-np-HCA** nonparametric agglomerative hierarchical cluster analysis with a proximity measure based on conditional item pair covariances 50f., 68

**IRT-pc2-ML** the two-parameter partial credit IRT model by means of ML 65, 68

**IRF** item response function 12ff., 31, 46, 49, 66, 236, 291

**IRT** item response theory vff., 1ff., 5, 7f., 12ff., 19ff., 27ff., 38ff., 49ff., 53, 55ff., 63ff., 69ff., 75f., 79ff., 84f., 91, 123, 161, 173, 190, 197f., 200, 202, 204, 206, 208, 210, 212, 214, 216, 218, 220, 222ff., 226, 228, 230, 232, 234, 236, 238, 240, 242, 244, 246, 248, 250ff., 258, 260, 293, 296, 363ff.

**IRT-2p** two-parameter IRT model 13ff., 28, 46, 54, 59, 63, 65, 259

**IRT-3p** the three-parameter IRT model 65

**IRT-grm** the graded response IRT model viii, 2, 15f., 19, 21f., 24f., 28, 59f., 65, 69f., 72f., 75, 79f., 84f., 88f., 91ff., 96ff., 101, 103, 105ff., 112ff., 132, 137ff., 143, 146, 149ff., 158ff., 173, 180f., 184ff., 195, 197f., 206, 212ff., 222f., 229, 234ff., 238, 242, 244f., 247ff., 252ff., 291, 294f., 300f., 303, 305f., 312f., 315, 317, 319, 321, 323, 325, 327, 329ff., 333, 335, 337, 339, 341, 343, 345, 347ff., 357, 360f., 363ff.

**IRT-mok** the nonparametric Mokken IRT model viif., 2, 16, 20, 24f., 69, 72f., 75, 79f., 84f., 89, 91ff., 97, 101, 104, 132f., 136ff., 141ff., 146f., 173f., 179ff., 184ff., 191ff., 197ff., 206, 212, 219ff., 229, 236ff., 242, 245ff., 252ff., 307, 352ff., 360f.

**ISRF** item-step response function 16, 219, 222, 245, 257

**LDR** leading digit rule 307f.

**LI** limited-information 23ff., 251, 369

**LV** latent variable viif., 1, 3, 5ff., 31, 35, 45ff., 67, 69ff., 75ff., 79ff., 84f., 87ff., 91ff., 101ff., 116, 126, 128f., 133, 136ff., 141ff., 171ff., 197ff., 206ff., 214f., 218f., 222f., 228ff., 238f., 242ff., 247ff., 285ff., 289ff., 299, 347, 356ff.

**MAD** median absolute deviation from the median 29, 34

**MANOVA** multivariate analysis of variance 90f., 103f., 111f., 116f., 120, 123, 125f., 133, 146ff., 156f., 161f., 165, 174f., 179, 356

**ML** maximum likelihood 10f., 19, 23f., 30, 36f., 61, 69, 199, 252, 255, 257ff.

**MLR** robust ML 19, 61, 69, 252, 257f.

**MML** marginal ML 16f., 19

**MSE** mean squared error 49, 51, 68

**NA** not available 199

**NIRT** nonparametric IRT 12, 16, 60, 64, 66

**NNFI** nonnormed fit index 37, 61f., 68

**PA** *Psychological Assessment* 27, 41

**PB** plain bias 86, 103, 111, 113, 117, 126, 128, 146ff., 156f., 300ff., 307, 311ff., 352ff.

**Q-Q** quantile-quantile 88, 103ff., 131, 171ff., 201ff., 225ff., 241, 298f., 309, 358f.

**RASJS** Revised Anticipated Sexual Jealousy Scale viif., 197, 223ff., 248, 365ff.

**RB** relative bias 86, 103, 111ff., 116f., 120, 123, 125, 128f., 133, 136, 140, 146ff., 155ff., 161f., 165, 174f., 179, 300ff., 307, 311ff., 352ff.

**RMR** root mean residuals 37, 61, 68

**RMSE** root mean squared error 47ff., 51, 53, 56, 64f., 68, 71, 86f., 90, 120, 125, 155, 300ff., 307, 311ff., 352ff.

**RMSEA** root mean squared error of approximation viii, 37, 61ff., 68, 73, 88ff., 94, 96ff., 103, 129ff., 141f., 146, 167ff., 188, 190, 192, 215ff., 235f., 244f., 255f., 260, 358f., 367

**SD** standard deviation 68, 105, 200, 224, 240, 300ff., 307, 311ff., 352ff.

**SRMR** standardized root mean residuals viii, 37, 61ff., 68, 73, 88ff., 94, 96, 98, 103f., 129f., 132, 141f., 146, 167ff., 173, 190f., 216ff., 235f., 244f., 256, 260, 291, 298, 367

**SSCP** sums of squares and cross-products 90

**TLI** Tucker-Lewis index 37, 63, 68, 216ff., 235f., 245, 367

**ULS** unweighted least squares 11, 61f., 70, 258

**WLS** weighted least squares 10f., 37, 61

**WLSMV** mean-and-variance adjusted weighted least squares 11, 37, 61, 69f., 105, 252, 258

# Appendix B

# Notation

The most important symbols used are given below.

$X_{is}$: an item score.

$X_{is}^{*}$: a latent, continuous item score.

$\omega$: a standardized population parameter.

$\omega^{\circ}$: an unstandardized population parameter.

$\hat{\omega}$: a standardized parameter estimator.

$\hat{\omega}^{\circ}$: an unstandardized parameter estimator.

$\theta_s$: an LV score.

$\lambda_i$: the loading of item $i$ on an LV.

$\tau_{ci}$: threshold $c$ of item $i$.

$\epsilon_i$: the residual part of item $i$.

$\alpha_i$: the discrimination parameter of item $i$.

$\beta_{ci}$: the $c$th (step-)difficulty parameter of item $i$.

$c : 1, 2, \ldots, C$: the $c$th item threshold.

$i : 1, 2, \ldots, I$: the $i$th item.

$s : 1, 2, \ldots, n$: the $s$th respondent.

$q : 1, 2, \ldots, Q$: the $q$th LV.

$\mathbf{\Sigma}$: an $(I \times I)$ population covariance matrix of the items.

$\mathbf{\Phi}$: a $(Q \times Q)$ LV covariance matrix.

$\mathbf{\Psi}$: an $(I \times I)$ error covariance matrix.

$R$: the number of replications.

$\varsigma$: the skewness of a distribution.

$\kappa$: the excess kurtosis of a distribution.

$\mathcal{N}(\mu, \sigma^2)$: a normal distribution with mean $\mu$ and variance $\sigma^2$.

$\mathcal{SN}(\iota)$: a skew-normal distribution, as a function of the normal density distribution, with a shape parameter $\iota$.

$\mathcal{SN}(p1, p2, \iota)$: a skew-normal distribution with a location, scale, and shape parameter, respectively.

$\mathcal{U}(a, b)$: a uniform distribution within the range $[a, b]$.

# Appendix C

# Setup of the Simulation Study

## C.1 Threshold Values

In the following tables the threshold values used in our simulation study are listed. In case of a normal LV, the standardized thresholds are independent of the loading values. For unstandardized thresholds or skew-normal LVs, threshold values are dependent on the loading values.

*Table C.1.* Standardized thresholds for a *normal* latent variable.

| Item distribution | $\tau_{i1}$ | $\tau_{i2}$ | $\tau_{i3}$ | $\tau_{i4}$ |
|---|---|---|---|---|
| Normal | −1.645 | −0.643 | 0.643 | 1.645 |
| Bimodal | −1.282 | −0.126 | 0.126 | 1.282 |
| Right-skewed | 0.346 | 0.800 | 1.221 | 1.622 |
| Left-skewed | −1.622 | −1.221 | −0.800 | −0.346 |

*Table C.2.* Unstandardized thresholds for a *normal* latent variable.

| Item loading | Item distribution | $\tau_{i1}$ | $\tau_{i2}$ | $\tau_{i3}$ | $\tau_{i4}$ |
|---|---|---|---|---|---|
| Strong | Normal | −2.741 | −1.072 | 1.072 | 2.741 |
| Medium | Normal | −1.899 | −0.743 | 0.743 | 1.899 |
| Weak | Normal | −1.724 | −0.674 | 0.674 | 1.724 |
| Strong | Bimodal | −2.136 | −0.209 | 0.209 | 2.136 |
| Medium | Bimodal | −1.480 | −0.145 | 0.145 | 1.480 |
| Weak | Bimodal | −1.343 | −0.132 | 0.132 | 1.343 |
| Strong | Right-skewed | 0.576 | 1.333 | 2.035 | 2.703 |
| Medium | Right-skewed | 0.399 | 0.924 | 1.410 | 1.872 |
| Weak | Right-skewed | 0.362 | 0.839 | 1.280 | 1.700 |
| Strong | Left-skewed | −2.703 | −2.035 | −1.333 | −0.576 |
| Medium | Left-skewed | −1.872 | −1.410 | −0.924 | −0.399 |
| Weak | Left-skewed | −1.700 | −1.280 | −0.839 | −0.362 |

*Table C.3.* Standardized thresholds for a *right skew-normal* latent variable.

| Item loading | Item distribution | $\tau_{i1}$ | $\tau_{i2}$ | $\tau_{i3}$ | $\tau_{i4}$ |
|---|---|---|---|---|---|
| Strong | Normal | −1.483 | −0.680 | 0.583 | 1.790 |
| Medium | Normal | −1.609 | −0.652 | 0.629 | 1.679 |
| Weak | Normal | −1.637 | −0.645 | 0.640 | 1.652 |
| Strong | Bimodal | −1.203 | −0.212 | 0.034 | 1.334 |
| Medium | Bimodal | −1.265 | −0.145 | 0.105 | 1.292 |
| Weak | Bimodal | −1.278 | −0.130 | 0.121 | 1.284 |
| Strong | Right-skewed | 0.260 | 0.760 | 1.260 | 1.760 |
| Medium | Right-skewed | 0.326 | 0.789 | 1.228 | 1.654 |
| Weak | Right-skewed | 0.342 | 0.798 | 1.223 | 1.628 |
| Strong | Left-skewed | −1.465 | −1.156 | −0.813 | −0.416 |
| Medium | Left-skewed | −1.587 | −1.208 | −0.804 | −0.362 |
| Weak | Left-skewed | −1.614 | −1.219 | −0.801 | −0.349 |

*Table C.4.* Unstandardized thresholds for a *right skew-normal* latent variable.

| Item loading | Item distribution | $\tau_{i1}$ | $\tau_{i2}$ | $\tau_{i3}$ | $\tau_{i4}$ |
|---|---|---|---|---|---|
| Strong | Normal | −2.471 | −1.133 | 0.971 | 2.983 |
| Medium | Normal | −1.858 | −0.753 | 0.726 | 1.939 |
| Weak | Normal | −1.716 | −0.677 | 0.671 | 1.732 |
| Strong | Bimodal | −2.005 | −0.353 | 0.056 | 2.223 |
| Medium | Bimodal | −1.461 | −0.167 | 0.121 | 1.492 |
| Weak | Bimodal | −1.340 | −0.136 | 0.127 | 1.346 |
| Strong | Right-skewed | 0.433 | 1.267 | 2.100 | 2.933 |
| Medium | Right-skewed | 0.377 | 0.912 | 1.418 | 1.910 |
| Weak | Right-skewed | 0.358 | 0.836 | 1.282 | 1.707 |
| Strong | Left-skewed | −2.442 | −1.926 | −1.355 | −0.694 |
| Medium | Left-skewed | −1.833 | −1.395 | −0.928 | −0.418 |
| Weak | Left-skewed | −1.692 | −1.277 | −0.840 | −0.366 |

## C.2 Illustration: From LV to Sum Score

To clarify the relation between $\theta_s$, $X^*_{is}$, $X_{is}$, and scale scores estimated by summing over item scores, an illustration of these relations is provided by presenting two normal item variables, one loading on a normal and one loading on a skew-normal LV.

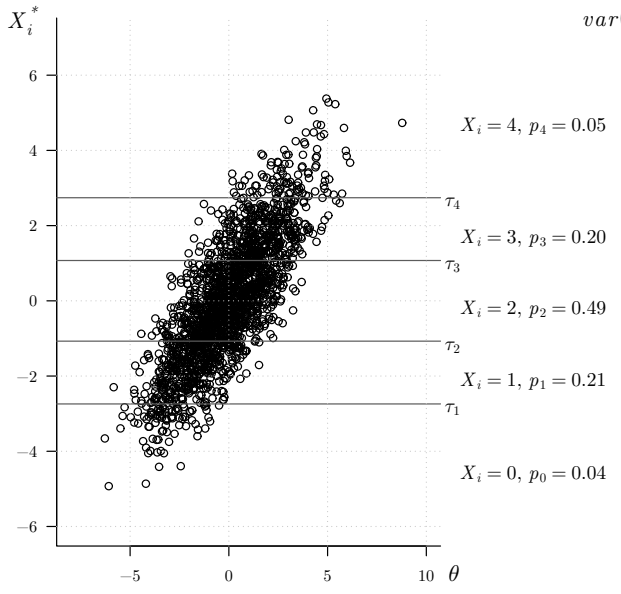
In Figure C.1 latent continuous item scores $X^*_{is}$ are plotted for a sample of LV scores $\theta$, in case of a normal LV distribution (upper left panel) and in case of a right skew-normal LV distribution (lower left panel). In both cases, the ordered categorical $X_{is}$ is approximately normal, which is accomplished by taking different sets of thresholds, marked by the horizontal lines. The corresponding $X_{is}$ values are indicated in the right margin of the plots in the left panel, as well as the corresponding proportion $p_i$ of respondents in these samples.

As explained, we can generate approximately normal items that load on a skew-normal LV, by choosing the thresholds carefully. When we create a scale of such items, the distribution of the sum score over items is, evidently, a sum of approximately normal variables. Because taking the sum over the item scores is the simplest way of estimating a respondent's LV score, sum scores are widely used for LV score approximation, regardless of the employed scaling model. Although one might expect the sum scores of normal items to be normally distributed as well, this is not the case in these data configurations of normal items loading on a skew-normal LV. In the following, we illustrate and explain this peculiarity.

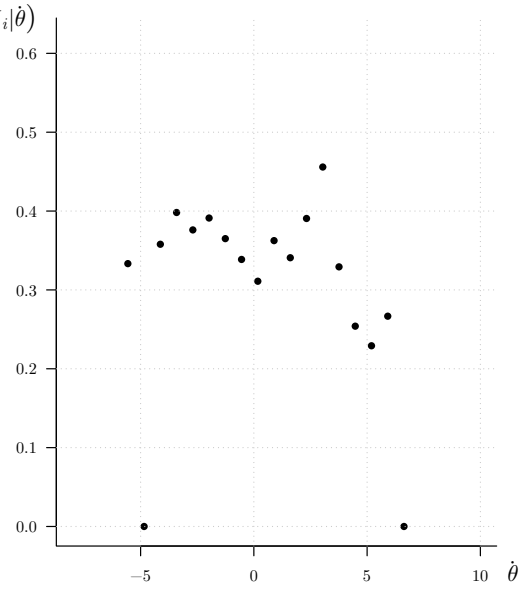
Why do sum scores of normal items follow different distributions depending on the LV distribution underlying the items? First, notice that in Figure C.1 the range of LV values differs between the plots. While the sample $\theta_s$ values from the normal LV distribution range between $-7$ and 7, those from the right skew-normal LV distribution are between $-4$ and 10. Second, observe the different shapes of the cloud of points in both plots. For the normal LV distribution, the cloud of points is ellipsoid, which is expected for bivariate normal variables and indicative of a linear relation between the $\theta_s$ values and the sum scores. For the skew-normal LV distribution, we observe a large variation in sum scores for the lower $\theta_s$ values. This is also apparent from the right panels of Figure C.1, where the variance of categorical $X_i$ conditional on $\theta_s$ in discretized form (denoted by $\dot{\theta}_s$) is depicted. The conditional variance of $X_i$ was computed by discretizing $\theta$ into 21 intervals and calculating for each interval the variance of corresponding $X_i$ values. The variance of the categorical $X_i$ decreases with increasing $\theta_s$ in case of a right skew-normal LV.

Going back to the true LV distributions, both constructed to have a mean of zero, we know that the median of our right skew-normal distribution equals $-0.39$. This means that half the $\theta_s$ values are between $-4$ and $-0.39$ and half are between $-0.39$ and 10. For the right skew-normal LV $X^*_{is}$ cannot be predicted well by $\theta_s$ for the lower half of the scale (below $-0.39$). For increasing $\theta_s$ values the prediction improves. This does not hold for the normal LV, where the quality of the prediction is rather constant over the range of $\theta_s$.

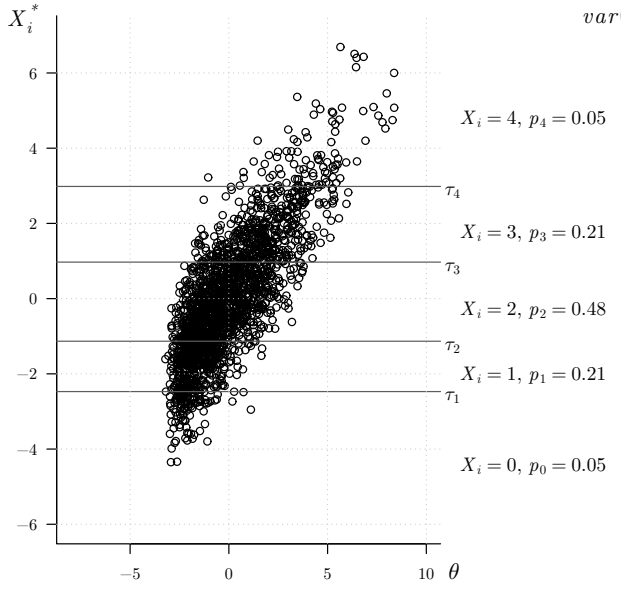
Since $\theta_s$ is constant for a respondent in a sample, the displayed relation holds for each arbitrary item in a scale of homogeneous item loadings. The variation on



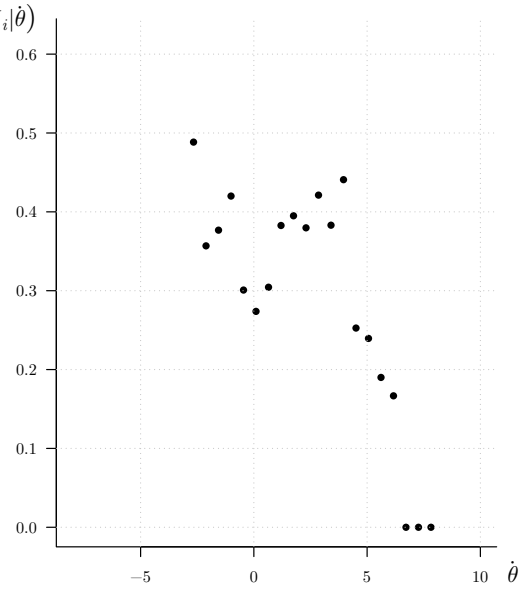

Sampled LV values $\theta$ from a normal LV distribution




Discretized LV values $\dot{\theta}$ from a normal LV distribution




Sampled LV values $\theta$ from a right skew-normal LV distribution




Discretized LV values $\dot{\theta}$ from a right skew-normal LV distribution

*Figure C.1.* Latent continuous item score $X_{is}^*$ for sampled $\theta$ values in case of a normal (upper panel) and right skew-normal (lower panel) LV distribution, as well as the variance of the corresponding ordered categorical $X_{is}$ conditional on a discretization of the sampled LV scores $\dot{\theta}$. $n = 2000$.

the vertical axis, conditional on the position at the horizontal axis, can therefore be viewed as the variation in $X_{is}^*$ of a single respondent.

So the error term in $X_{is}^* = \lambda_i \theta_s + \epsilon_{is}$ is of much more influence in determining the observed categorical $X_{is}$ value in the lower half than in the upper half of the $\theta$ distribution in case of a right skew-normal underlying LV. This causes an increased variance in item scores for the lower half of the distribution, resulting in the absence of a lower tail in the sum score distribution. For the upper half of the distribution the opposite holds. As many $\theta_s$ values are much larger than the upper threshold, the error term is of little influence and large $\theta_s$ values lead to maximum item scores and thus to a maximum sum score, resulting in a distinctive upper tail in the sum score distribution.

## C.3 Model-Implied Covariance Matrix for FA-poly and IRT-grm

In order to compute the SRMR and the $\chi^2_{YB}$, the model-implied covariance matrix $\boldsymbol{\Sigma}$ of the observed categorical items $\mathbf{X}$ is required. Maydeu-Olivares et al. (2011) show in their appendix that the variance and covariance of the observed categorical item variables $X_i$ and $X_j$ are, respectively,

$$\sigma_i^2 = \int_{-\infty}^{+\infty} \left( \sum_{c=1}^{C-1} (2c-1)\, \text{IRF}_{i,c} \right) f(\theta) d(\theta) - \mu_i^2, \tag{C.1}$$

$$\sigma_{ij} = \left( \int_{-\infty}^{+\infty} \left[ \sum_{d=1}^{C-1} \text{IRF}_{j,d} \right] \left[ \sum_{c=1}^{C-1} \text{IRF}_{i,c} \right] f(\theta) d(\theta) \right) - \mu_i \mu_j, \tag{C.2}$$

where

$$\mu_i = \int_{-\infty}^{+\infty} \left[ \sum_{c=1}^{C-1} \text{IRF}_{i,c} \right] f(\theta) d(\theta), \tag{C.3}$$

and $\text{IRF}_{i,c}$ is the logistic item response function for item $i$ and category $c$, given in Equation 1.21.

For the FA-poly model, the model-implied covariance matrix is computed by applying the normal-ogive IRF (see Equation 1.20) to Equations C.1 to C.3.

## C.4 Illustration of Data Generation

We give an illustration of the data generation using our R code (which is available from the author upon request). It concerns a small data set with scores of $n = 5$ respondents on four five-category items: two approximately normally distributed, one left-skewed, and one right-skewed, loading 0.80, 0.50, 0.80 and 0.50, respectively, on

one normal LV with a variance of 4. Unstandardized error variances are set to 1. The seed used for (pseudo)-random number generation here is 491614761.

The standardized input loadings and thresholds are specified as follows:

$$\boldsymbol{\lambda} = (\ 0.80 \quad 0.50 \quad 0.80 \quad 0.50\ )',$$

$$\boldsymbol{\tau} = \begin{bmatrix} -1.645 & -1.645 & -1.622 & 0.346 \\ -0.643 & -0.643 & -1.221 & 0.800 \\ 0.643 & 0.643 & -0.800 & 1.221 \\ 1.645 & 1.645 & -0.346 & 1.622 \end{bmatrix}.$$

To compute the unstandardized loadings and thresholds, $\sigma^{\circ}_{i*}$ is determined for each item $i$ using Equations 4.13 and 4.14,

$$\boldsymbol{\sigma}^{\circ}_{*} = (\ 2.78 \quad 1.33 \quad 2.78 \quad 1.33\ )'.$$

Unstandardized loadings and thresholds can now be computed using Equations 4.12 and 4.15, respectively,

$$\boldsymbol{\lambda}^{\circ} = (\ 0.667 \quad 0.289 \quad 0.667 \quad 0.289\ )',$$

$$\boldsymbol{\tau}^{\circ} = \begin{bmatrix} -2.741 & -1.899 & -2.703 & 0.399 \\ -1.072 & -0.743 & -2.035 & 0.924 \\ 1.072 & 0.743 & -1.333 & 1.410 \\ 2.741 & 1.899 & -0.576 & 1.872 \end{bmatrix}.$$

Using Equation 4.3, the population covariance matrix of the latent item scores $\mathbf{X}^{*}$ is thus

$$\boldsymbol{\Sigma}^{\circ}_{*} = \begin{bmatrix} 2.78 & 0.77 & 1.78 & 0.77 \\ 0.77 & 1.33 & 0.77 & 0.33 \\ 1.78 & 0.77 & 2.78 & 0.77 \\ 0.77 & 0.33 & 0.77 & 1.33 \end{bmatrix}.$$

The LV and error scores are drawn from a $\mathcal{N}(0, 4)$ normal and $\mathcal{N}(0, 1)$ standard normal distribution, respectively, giving

$$\boldsymbol{\theta}^{\circ} = (\ -1.149 \quad -1.465 \quad 1.826 \quad -0.048 \quad 2.699\ )',$$

the unstandardized latent scores of five respondents, and the unstandardized error matrix corresponding to Equation 1.7

$$\mathbf{E}^{\circ} = \begin{bmatrix} 0.233 & -1.007 & 1.911 & 0.382 \\ -1.797 & 0.348 & -0.343 & 1.016 \\ 1.237 & 0.341 & 0.768 & 0.305 \\ 2.179 & 0.539 & 0.568 & 1.601 \\ 0.739 & 0.681 & -0.543 & 0.649 \end{bmatrix}.$$

Given this input, $\mathbf{X}^*$ and $\mathbf{X}$ can be computed using Equation 1.7,

$$\mathbf{X}^* = \begin{bmatrix} -0.533 & -1.339 & 1.145 & 0.051 \\ -2.773 & -0.075 & -1.320 & 0.593 \\ 2.455 & 0.868 & 1.985 & 0.832 \\ 2.147 & 0.525 & 0.536 & 1.587 \\ 2.538 & 1.461 & 1.256 & 1.428 \end{bmatrix},$$

$$\mathbf{X} = \begin{bmatrix} 2 & 1 & 4 & 0 \\ 0 & 2 & 3 & 1 \\ 3 & 3 & 4 & 1 \\ 3 & 2 & 4 & 3 \\ 3 & 3 & 4 & 3 \end{bmatrix}.$$

In addition, the unstandardized IRT parameters $\alpha^\circ$ and $\beta^\circ$ are computed by applying Equations 1.43 and 4.5, respectively, on unstandardized $\lambda^\circ$, $\psi^\circ$, and $\tau^\circ$,

$$\boldsymbol{\alpha}^\circ = \begin{pmatrix} 1.135 & 0.491 & 1.135 & 0.491 \end{pmatrix}',$$

$$\boldsymbol{\beta}^\circ = \begin{bmatrix} -4.112 & -6.579 & -4.054 & 1.383 \\ -1.608 & -2.573 & -3.053 & 3.200 \\ 1.608 & 2.573 & -2.000 & 4.885 \\ 4.112 & 6.579 & -0.864 & 6.486 \end{bmatrix}.$$

Standardized $\alpha$ and $\beta$ are computed by applying Equations 4.10 and 4.11,

$$\boldsymbol{\alpha} = \begin{pmatrix} 2.269 & 0.983 & 2.269 & 0.983 \end{pmatrix}',$$

$$\boldsymbol{\beta} = \begin{bmatrix} -2.056 & -3.290 & -2.027 & 0.692 \\ -0.804 & -1.287 & -1.527 & 1.600 \\ 0.804 & 1.287 & -1.000 & 2.442 \\ 2.056 & 3.290 & -0.432 & 3.243 \end{bmatrix}.$$

Population Loevinger's $H$ values are computed by first calculating the bivariate probability tables of item pairs using Equation 4.7. We give these only for item pair (1,2),

$$P_{1,2} = \begin{bmatrix} 0.009 & 0.019 & 0.018 & 0.003 & 0.000 \\ 0.019 & 0.066 & 0.099 & 0.023 & 0.003 \\ 0.018 & 0.099 & 0.246 & 0.099 & 0.018 \\ 0.003 & 0.023 & 0.099 & 0.066 & 0.019 \\ 0.000 & 0.003 & 0.018 & 0.019 & 0.009 \end{bmatrix}.$$

From the bivariate tables we can calculate $F_{ij}$ and $E_{ij}$ by use of our own implementation of I. W. Molenaar (1991), from which $H$ values are computed using Equa-

tions 1.36 to 1.38,

$$H_{ij} = \begin{bmatrix} 1.000 & 0.357 & 0.631 & 0.394 \\ 0.357 & 1.000 & 0.394 & 0.246 \\ 0.631 & 0.394 & 1.000 & 0.587 \\ 0.394 & 0.246 & 0.587 & 1.000 \end{bmatrix},$$

$$H_i = \begin{pmatrix} 0.461 & 0.332 & 0.531 & 0.386 \end{pmatrix}',$$

$$H_{scale} = 0.426.$$

After samples of data have been generated, estimation of the parametric models FA-lin, FA-poly, and IRT-grm is performed by running MPLUS. The MPLUS input files used for each of the respective models are presented next. Each analysis is applied to the same data set.

Note that the starting values for the thresholds, which are equal to the true model parameters, from the IRT-grm file cannot be found in the parameter sets just presented. This is due to the fact that in MPLUS IRT-grm thresholds are put on the logit scale. Hence the values are obtained by multiplying the threshold values by 1.702.

Listing C.1: FA-lin MPLUS input file

```
TITLE:
  FA-lin analysis
data:
  file = mydata.dat;
  type = montecarlo;
variable:
  names = v1-v4;
model:
  f by v1*0.667 v2*0.289 v3*0.667 v4*0.289;
  f@4;
output:
  Stdyx;
  tech9;
savedata:
  results = myresults.dat;
```

Listing C.2: FA-poly MPLUS input file

```
TITLE:
  FA-poly analysis
data:
  file = mydata.dat;
  type = montecarlo;
variable:
  names = v1-v4;
  categorical = v1-v4;
analysis:
  parameterization = theta;
model:
  f by v1*0.667 v2*0.289 v3*0.667 v4*0.289;
  f@4;
  [v1$1*-2.741]; [v1$2*-1.072]; [v1$3*1.072]; [v1$4*2.741];
  [v1$1*-1.899]; [v1$2*-0.743]; [v1$3*0.743]; [v1$4*1.899];
  [v1$1*-2.741]; [v1$2*-1.072]; [v1$3*1.072]; [v1$4*2.741];
  [v1$1*-1.899]; [v1$2*-0.743]; [v1$3*0.743]; [v1$4*1.899];
output:
  Stdyx;
  tech9;
savedata:
  results = myresults.dat;
```

Listing C.3: IRT-grm MPLUS input file

```
TITLE:
  IRT-grm analysis
data:
  file = mydata.dat;
  type = montecarlo;
variable:
  names = v1-v4;
  categorical = v1-v4;
analysis:
  estimator = MLR;
model:
  f by v1*1.135 v2*0.491 v3*1.135 v4*0.491;
  f@4;
  [v1$1*-4.666]; [v1$2*-1.825]; [v1$3*1.825]; [v1$4*4.666];
  [v1$1*-3.233]; [v1$2*-1.264]; [v1$3*1.264]; [v1$4*3.233];
  [v1$1*-4.666]; [v1$2*-1.825]; [v1$3*1.825]; [v1$4*4.666];
  [v1$1*-3.233]; [v1$2*-1.264]; [v1$3*1.264]; [v1$4*3.233];
output:
  Stdyx;
  tech8 tech9;
savedata:
```

```
results = myresults.dat;
```

## C.5 FA and Corresponding IRT Parameters

*Table C.5.* Standardized and unstandardized FA and corresponding IRT parameter values for approximately normal items and a normal latent variable.

| $\lambda$ | $\tau$ | $\alpha$ | $\beta$ | $\lambda^\circ$ | $\tau^\circ$ | $\alpha^\circ$ | $\beta^\circ$ |
|---|---|---|---|---|---|---|---|
| 0.8 | −1.645 | 2.269 | −2.056 | 0.667 | −2.741 | 1.135 | −4.112 |
| | −0.643 | | −0.804 | | −1.072 | | −1.608 |
| | 0.643 | | 0.804 | | 1.072 | | 1.608 |
| | 1.645 | | 2.056 | | 2.741 | | 4.112 |
| 0.5 | −1.645 | 0.983 | −3.290 | 0.289 | −1.899 | 0.491 | −6.579 |
| | −0.643 | | −1.287 | | −0.743 | | −2.573 |
| | 0.643 | | 1.287 | | 0.743 | | 2.573 |
| | 1.645 | | 3.290 | | 1.899 | | 6.579 |
| 0.3 | −1.645 | 0.535 | −5.483 | 0.157 | −1.724 | 0.268 | −10.966 |
| | −0.643 | | −2.144 | | −0.674 | | −4.289 |
| | 0.643 | | 2.144 | | 0.674 | | 4.289 |
| | 1.645 | | 5.483 | | 1.724 | | 10.966 |

# Appendix D

# Simulation Study: Normal Configurations

## D.1 Seeds for Data Generation

*Table D.1.* Seeds for data generation for each cell of the design.

| Cell | Seed |
|---|---|
| nNS2 | 277223 |
| nNM2 | 962045 |
| rnNS2 | 385134 |
| lnNS2 | 587175 |
| lrnNS2 | 650454 |
| bnNS2 | 745967 |
| nRS2 | 505530 |
| rnRS2 | 397921 |
| lnRS2 | 214507 |
| lrnRS2 | 23312 |
| bnRS2 | 346104 |
| nNS6 | 142222 |
| nNM6 | 995007 |
| rnNS6 | 895330 |
| lnNS6 | 614782 |
| lrnNS6 | 998090 |
| bnNS6 | 726535 |
| nRS6 | 781900 |
| rnRS6 | 518507 |
| lnRS6 | 404851 |
| lrnRS6 | 611710 |
| bnRS6 | 394138 |

## D.2 Distribution of SRMR Fit Statistic

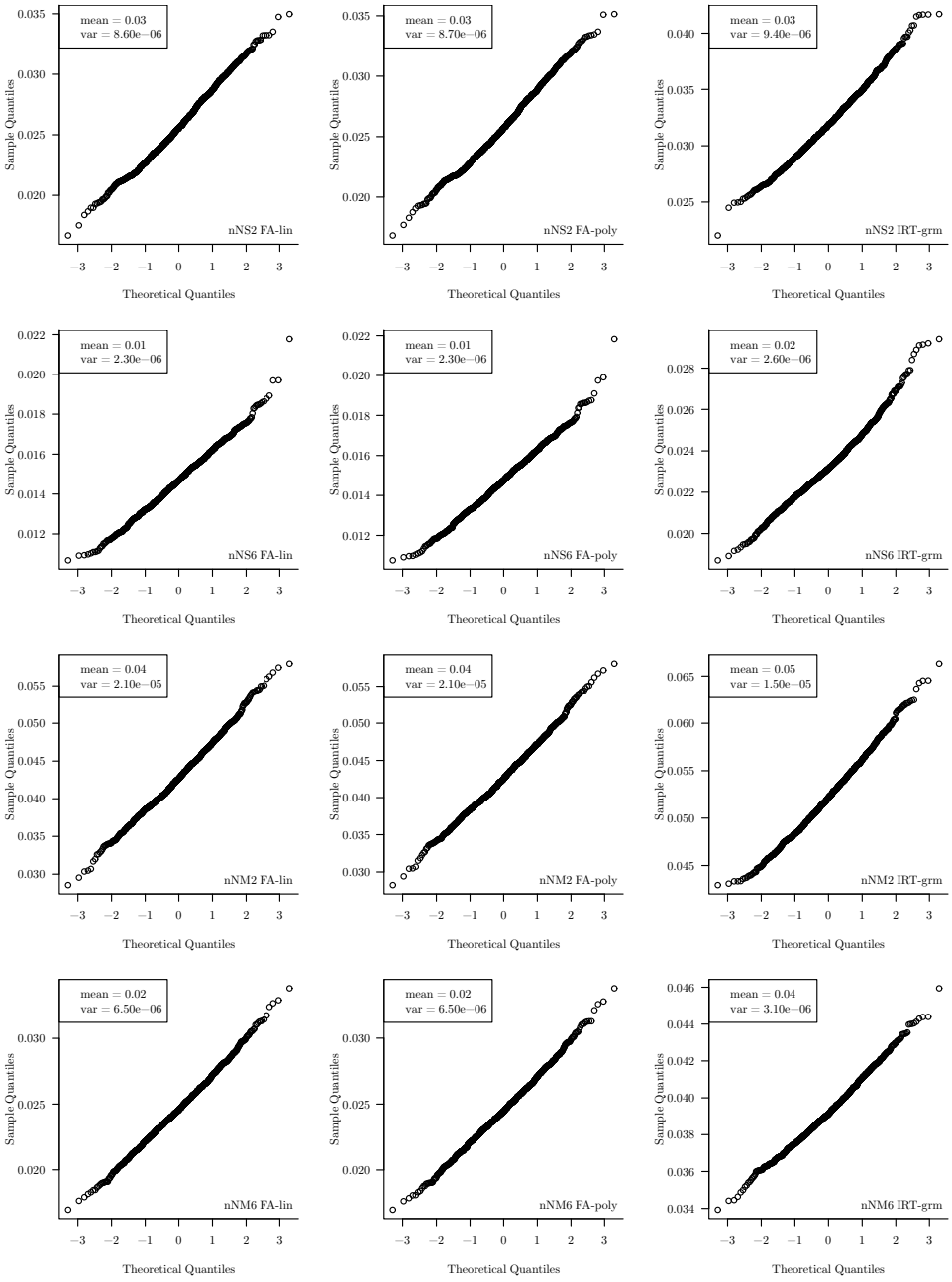



*Figure D.1.* Normal Q-Q plots for SRMR fit statistic for Cells nNS2, nNS6, nNM2, and nNM6 and each model. $R = 1000$.

## D.3 Distribution of Kendall's $\tau_a$ for LV scores

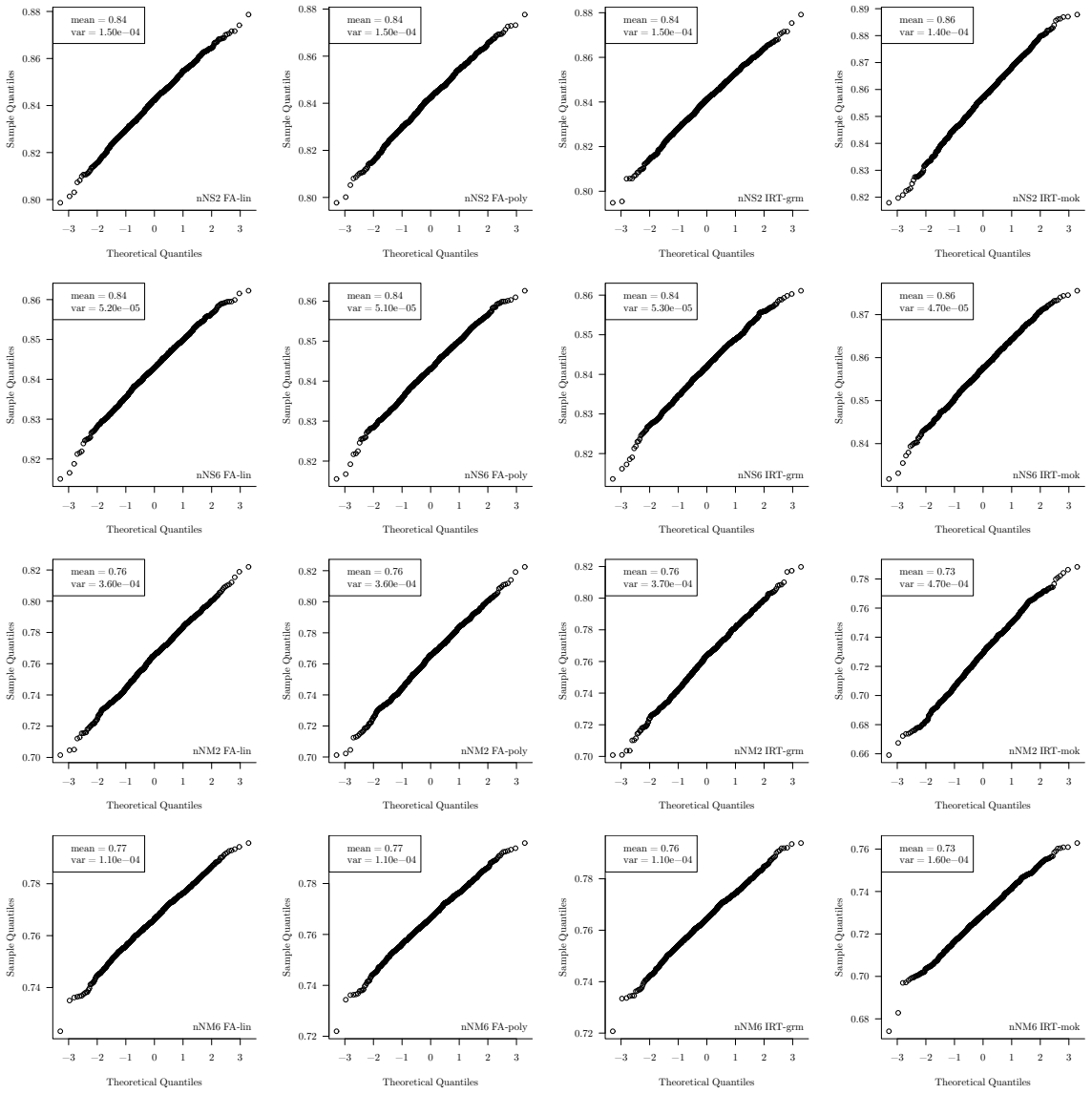



*Figure D.2.* Normal Q-Q plots for Kendall's $\tau_a$ association between true and estimated LV scores for Cells nNS2, nNS6, nNM2, and nNM6 and each model. $R = 1000$.

# D.4 Tables of Parameter and Standard Error Estimates

## D.4.1 Average Parameter and Standard Error Results

Parameter estimates are averaged over parameters with equal loading values (strong, medium, or weak). The coverage rate of the parameter's 95%-confidence interval is presented under 95%-cov.

*Table D.2.* FA-lin standardized parameters of interest and corresponding standard errors in Cell nNS2 averaged over items with equal population value/loading. $n = 200$; $R = 1000$.

| | $\omega$ | PB($\hat{\omega}$) | RB($\hat{\omega}$) | RMSE($\hat{\omega}$) | SD($\hat{\omega}$) | RB($\hat{se}$) | RMSE($\hat{se}$) | 95%-cov. |
|---|---|---|---|---|---|---|---|---|
| $\lambda_{(1–12)}$ | 0.800/strong | −0.045 | −0.056 | 0.056 | 0.033 | −0.002 | 0.004 | 0.780 |

*Table D.3.* FA-poly standardized parameters of interest and corresponding standard errors in Cell nNS2 averaged over items with equal population value/loading. $n = 200$; $R = 1000$.

| | $\omega$ | PB($\hat{\omega}$) | RB($\hat{\omega}$) | RMSE($\hat{\omega}$) | SD($\hat{\omega}$) | RB($\hat{se}$) | RMSE($\hat{se}$) | 95%-cov. |
|---|---|---|---|---|---|---|---|---|
| $\lambda_{(1–12)}$ | 0.800/strong | 0.003 | 0.003 | 0.033 | 0.032 | −0.052 | 0.004 | 0.927 |
| $\tau_{(1–12).1}$ | −1.645/strong | −0.023 | 0.014 | 0.163 | 0.161 | −0.044 | 0.020 | 0.949 |
| $\tau_{(1–12).2}$ | −0.643/strong | −0.002 | 0.004 | 0.096 | 0.096 | −0.001 | 0.003 | 0.948 |
| $\tau_{(1–12).3}$ | 0.643/strong | 0.004 | 0.006 | 0.096 | 0.096 | −0.003 | 0.003 | 0.947 |
| $\tau_{(1–12).4}$ | 1.645/strong | 0.020 | 0.012 | 0.161 | 0.160 | −0.040 | 0.020 | 0.947 |
| $\alpha_{(1–12)}$ | 2.269/strong | 0.049 | 0.022 | 0.270 | 0.266 | −0.054 | 0.033 | 0.944 |
| $\beta_{(1–12).1}$ | −2.056/strong | −0.028 | 0.013 | 0.239 | 0.237 | | | |
| $\beta_{(1–12).2}$ | −0.804/strong | −0.002 | 0.003 | 0.129 | 0.129 | | | |
| $\beta_{(1–12).3}$ | 0.804/strong | 0.004 | 0.005 | 0.129 | 0.129 | | | |
| $\beta_{(1–12).4}$ | 2.056/strong | 0.024 | 0.012 | 0.238 | 0.237 | | | |

*Table D.4.* IRT-grm standardized parameters of interest and corresponding standard errors in Cell nNS2 averaged over items with equal population value/loading. $n = 200$; $R = 1000$.

| | $\omega$ | PB($\hat{\omega}$) | RB($\hat{\omega}$) | RMSE($\hat{\omega}$) | SD($\hat{\omega}$) | RB($\hat{se}$) | RMSE($\hat{se}$) | 95%-cov. |
|---|---|---|---|---|---|---|---|---|
| $\lambda_{(1–12)}$ | 0.800/strong | −0.003 | −0.004 | 0.034 | 0.034 | −0.010 | 0.004 | 0.946 |
| $\tau_{(1–12).1}$ | −1.645/strong | −0.016 | 0.010 | 0.165 | 0.164 | −0.054 | 0.024 | 0.952 |
| $\tau_{(1–12).2}$ | −0.643/strong | 0.009 | −0.014 | 0.094 | 0.094 | −0.006 | 0.004 | 0.948 |
| $\tau_{(1–12).3}$ | 0.643/strong | −0.008 | −0.012 | 0.095 | 0.094 | −0.012 | 0.004 | 0.945 |
| $\tau_{(1–12).4}$ | 1.645/strong | 0.014 | 0.008 | 0.165 | 0.164 | −0.056 | 0.024 | 0.948 |
| $\alpha_{(1–12)}$ | 2.269/strong | 0.004 | 0.002 | 0.268 | 0.268 | −0.012 | 0.033 | 0.946 |
| $\beta_{(1–12).1}$ | −2.056/strong | −0.034 | 0.017 | 0.246 | 0.244 | | | |
| $\beta_{(1–12).2}$ | −0.804/strong | 0.006 | −0.007 | 0.126 | 0.126 | | | |
| $\beta_{(1–12).3}$ | 0.804/strong | −0.005 | −0.006 | 0.126 | 0.126 | | | |
| $\beta_{(1–12).4}$ | 2.056/strong | 0.031 | 0.015 | 0.246 | 0.244 | | | |

*Table D.5.* FA-lin standardized parameters of interest and corresponding standard errors in Cell nNS6 averaged over items with equal population value/loading. $n = 600$; $R = 1000$.

| | $\omega$ | PB($\hat{\omega}$) | RB($\hat{\omega}$) | RMSE($\hat{\omega}$) | SD($\hat{\omega}$) | RB($\hat{se}$) | RMSE($\hat{se}$) | 95%-cov. |
|---|---|---|---|---|---|---|---|---|
| $\lambda_{(1–12)}$ | 0.800/strong | −0.044 | −0.055 | 0.048 | 0.019 | 0.003 | 0.001 | 0.353 |

*Table D.6.* FA-poly standardized parameters of interest and corresponding standard errors in Cell nNS6 averaged over items with equal population value/loading. $n = 600$; $R = 1000$.

| | $\omega$ | PB($\hat{\omega}$) | RB($\hat{\omega}$) | RMSE($\hat{\omega}$) | SD($\hat{\omega}$) | RB($\hat{se}$) | RMSE($\hat{se}$) | 95%-cov. |
|---|---|---|---|---|---|---|---|---|
| $\lambda_{(1–12)}$ | 0.800/strong | 0.001 | 0.002 | 0.019 | 0.019 | −0.013 | 0.001 | 0.943 |
| $\tau_{(1–12).1}$ | −1.645/strong | −0.004 | 0.003 | 0.088 | 0.087 | −0.007 | 0.005 | 0.951 |
| $\tau_{(1–12).2}$ | −0.643/strong | −0.000 | 0.000 | 0.056 | 0.056 | −0.014 | 0.001 | 0.944 |
| $\tau_{(1–12).3}$ | 0.643/strong | 0.000 | 0.001 | 0.056 | 0.056 | −0.014 | 0.002 | 0.946 |
| $\tau_{(1–12).4}$ | 1.645/strong | 0.004 | 0.003 | 0.088 | 0.087 | −0.006 | 0.006 | 0.950 |
| $\alpha_{(1–12)}$ | 2.269/strong | 0.019 | 0.008 | 0.149 | 0.148 | −0.009 | 0.010 | 0.949 |
| $\beta_{(1–12).1}$ | −2.056/strong | −0.004 | 0.002 | 0.131 | 0.131 | | | |
| $\beta_{(1–12).2}$ | −0.804/strong | 0.000 | −0.000 | 0.075 | 0.075 | | | |
| $\beta_{(1–12).3}$ | 0.804/strong | −0.000 | −0.000 | 0.075 | 0.075 | | | |
| $\beta_{(1–12).4}$ | 2.056/strong | 0.004 | 0.002 | 0.130 | 0.130 | | | |

*Table D.7.* IRT-grm standardized parameters of interest and corresponding standard errors in Cell nNS6 averaged over items with equal population value/loading. $n = 600$; $R = 1000$.

| | $\omega$ | PB($\hat{\omega}$) | RB($\hat{\omega}$) | RMSE($\hat{\omega}$) | SD($\hat{\omega}$) | RB($\hat{se}$) | RMSE($\hat{se}$) | 95%-cov. |
|---|---|---|---|---|---|---|---|---|
| $\lambda_{(1–12)}$ | 0.800/strong | −0.004 | −0.005 | 0.020 | 0.019 | 0.003 | 0.002 | 0.952 |
| $\tau_{(1–12).1}$ | −1.645/strong | 0.003 | −0.002 | 0.088 | 0.088 | −0.008 | 0.006 | 0.948 |
| $\tau_{(1–12).2}$ | −0.643/strong | 0.010 | −0.016 | 0.055 | 0.054 | −0.008 | 0.001 | 0.941 |
| $\tau_{(1–12).3}$ | 0.643/strong | −0.010 | −0.016 | 0.055 | 0.054 | −0.005 | 0.002 | 0.945 |
| $\tau_{(1–12).4}$ | 1.645/strong | −0.002 | −0.001 | 0.088 | 0.088 | −0.009 | 0.006 | 0.947 |
| $\alpha_{(1–12)}$ | 2.269/strong | −0.020 | −0.009 | 0.150 | 0.149 | 0.007 | 0.011 | 0.944 |
| $\beta_{(1–12).1}$ | −2.056/strong | −0.008 | 0.004 | 0.133 | 0.132 | | | |
| $\beta_{(1–12).2}$ | −0.804/strong | 0.009 | −0.011 | 0.073 | 0.073 | | | |
| $\beta_{(1–12).3}$ | 0.804/strong | −0.009 | −0.011 | 0.073 | 0.072 | | | |
| $\beta_{(1–12).4}$ | 2.056/strong | 0.009 | 0.004 | 0.132 | 0.132 | | | |

*Table D.8.* FA-lin standardized parameters of interest and corresponding standard errors in Cell nNM2 averaged over items with equal population value/loading. $n = 200$; $R = 1000$.

| | $\omega$ | PB($\hat{\omega}$) | RB($\hat{\omega}$) | RMSE($\hat{\omega}$) | SD($\hat{\omega}$) | RB($\hat{se}$) | RMSE($\hat{se}$) | 95%-cov. |
|---|---|---|---|---|---|---|---|---|
| $\lambda_{(1–4)}$ | 0.800/strong | −0.045 | −0.057 | 0.058 | 0.037 | 0.008 | 0.004 | 0.830 |
| $\lambda_{(5–8)}$ | 0.500/medium | −0.029 | −0.058 | 0.068 | 0.061 | −0.014 | 0.005 | 0.935 |
| $\lambda_{(9–12)}$ | 0.300/weak | −0.019 | −0.064 | 0.072 | 0.069 | 0.009 | 0.003 | 0.941 |

*Table D.9.* FA-poly standardized parameters of interest and corresponding standard errors in Cell nNM2 averaged over items with equal population value/loading. $n = 200$; $R = 1000$.

| | $\omega$ | PB($\hat{\omega}$) | RB($\hat{\omega}$) | RMSE($\hat{\omega}$) | SD($\hat{\omega}$) | RB($\hat{se}$) | RMSE($\hat{se}$) | 95%-cov. |
|---|---|---|---|---|---|---|---|---|
| $\lambda_{(1–4)}$ | 0.800/strong | 0.001 | 0.001 | 0.037 | 0.037 | −0.024 | 0.005 | 0.936 |
| $\lambda_{(5–8)}$ | 0.500/medium | 0.002 | 0.005 | 0.064 | 0.064 | −0.075 | 0.007 | 0.921 |
| $\lambda_{(9–12)}$ | 0.300/weak | 0.000 | 0.001 | 0.074 | 0.074 | −0.060 | 0.006 | 0.927 |
| $\tau_{(1–4).1}$ | −1.645/strong | −0.016 | 0.010 | 0.159 | 0.158 | −0.030 | 0.019 | 0.950 |
| $\tau_{(1–4).2}$ | −0.643/strong | 0.001 | −0.002 | 0.097 | 0.097 | −0.015 | 0.003 | 0.946 |
| $\tau_{(1–4).3}$ | 0.643/strong | 0.004 | 0.006 | 0.097 | 0.097 | −0.013 | 0.003 | 0.945 |
| $\tau_{(1–4).4}$ | 1.645/strong | 0.024 | 0.015 | 0.161 | 0.159 | −0.033 | 0.019 | 0.950 |
| $\tau_{(5–8).1}$ | −1.645/medium | −0.018 | 0.011 | 0.154 | 0.153 | −0.001 | 0.018 | 0.953 |
| $\tau_{(5–8).2}$ | −0.643/medium | −0.003 | 0.005 | 0.098 | 0.098 | −0.020 | 0.004 | 0.942 |
| $\tau_{(5–8).3}$ | 0.643/medium | 0.004 | 0.007 | 0.097 | 0.096 | −0.005 | 0.002 | 0.948 |
| $\tau_{(5–8).4}$ | 1.645/medium | 0.021 | 0.013 | 0.156 | 0.155 | −0.008 | 0.018 | 0.952 |
| $\tau_{(9–12).1}$ | −1.645/weak | −0.021 | 0.013 | 0.159 | 0.158 | −0.028 | 0.019 | 0.948 |
| $\tau_{(9–12).2}$ | −0.643/weak | −0.004 | 0.006 | 0.097 | 0.097 | −0.007 | 0.003 | 0.947 |
| $\tau_{(9–12).3}$ | 0.643/weak | 0.004 | 0.006 | 0.097 | 0.097 | −0.007 | 0.003 | 0.951 |
| $\tau_{(9–12).4}$ | 1.645/weak | 0.020 | 0.012 | 0.162 | 0.160 | −0.044 | 0.020 | 0.946 |
| $\alpha_{(1–4)}$ | 2.269/strong | 0.044 | 0.020 | 0.304 | 0.300 | −0.023 | 0.050 | 0.949 |
| $\alpha_{(5–8)}$ | 0.983/medium | 0.017 | 0.018 | 0.173 | 0.172 | −0.077 | 0.018 | 0.934 |
| $\alpha_{(9–12)}$ | 0.535/weak | 0.006 | 0.012 | 0.147 | 0.147 | −0.059 | 0.011 | 0.936 |
| $\beta_{(1–4).1}$ | −2.056/strong | −0.024 | 0.012 | 0.242 | 0.241 | | | |
| $\beta_{(1–4).2}$ | −0.804/strong | −0.000 | 0.000 | 0.133 | 0.133 | | | |
| $\beta_{(1–4).3}$ | 0.804/strong | 0.007 | 0.008 | 0.132 | 0.132 | | | |
| $\beta_{(1–4).4}$ | 2.056/strong | 0.035 | 0.017 | 0.244 | 0.242 | | | |
| $\beta_{(5–8).1}$ | −3.290/medium | −0.086 | 0.026 | 0.617 | 0.611 | | | |
| $\beta_{(5–8).2}$ | −1.287/medium | −0.025 | 0.019 | 0.282 | 0.281 | | | |
| $\beta_{(5–8).3}$ | 1.287/medium | 0.028 | 0.022 | 0.283 | 0.281 | | | |
| $\beta_{(5–8).4}$ | 3.290/medium | 0.091 | 0.028 | 0.614 | 0.607 | | | |
| $\beta_{(9–12).1}$ | −5.483/weak | −0.550 | 0.100 | 2.494 | 2.433 | | | |
| $\beta_{(9–12).2}$ | −2.144/weak | −0.196 | 0.092 | 0.993 | 0.974 | | | |
| $\beta_{(9–12).3}$ | 2.144/weak | 0.197 | 0.092 | 0.995 | 0.975 | | | |
| $\beta_{(9–12).4}$ | 5.483/weak | 0.545 | 0.099 | 2.516 | 2.457 | | | |

*Table D.10.* IRT-grm standardized parameters of interest and corresponding standard errors in Cell nNM2 averaged over items with equal population value/loading. $n = 200$; $R = 1000$.

| | $\omega$ | PB($\hat{\omega}$) | RB($\hat{\omega}$) | RMSE($\hat{\omega}$) | SD($\hat{\omega}$) | RB($\hat{se}$) | RMSE($\hat{se}$) | 95%-cov. |
|---|---|---|---|---|---|---|---|---|
| $\lambda_{(1–4)}$ | 0.800/strong | −0.003 | −0.004 | 0.039 | 0.038 | 0.016 | 0.005 | 0.948 |
| $\lambda_{(5–8)}$ | 0.500/medium | −0.013 | −0.026 | 0.066 | 0.065 | −0.005 | 0.007 | 0.945 |
| $\lambda_{(9–12)}$ | 0.300/weak | −0.013 | −0.045 | 0.074 | 0.073 | 0.020 | 0.007 | 0.944 |
| $\tau_{(1–4).1}$ | −1.645/strong | −0.011 | 0.007 | 0.161 | 0.161 | −0.038 | 0.022 | 0.949 |
| $\tau_{(1–4).2}$ | −0.643/strong | 0.013 | −0.020 | 0.096 | 0.095 | −0.015 | 0.003 | 0.942 |
| $\tau_{(1–4).3}$ | 0.643/strong | −0.007 | −0.012 | 0.095 | 0.095 | −0.012 | 0.004 | 0.945 |
| $\tau_{(1–4).4}$ | 1.645/strong | 0.018 | 0.011 | 0.164 | 0.163 | −0.045 | 0.023 | 0.951 |
| $\tau_{(5–8).1}$ | −1.645/medium | −0.003 | 0.002 | 0.173 | 0.173 | −0.007 | 0.028 | 0.949 |
| $\tau_{(5–8).2}$ | −0.643/medium | 0.038 | −0.058 | 0.101 | 0.094 | −0.023 | 0.004 | 0.920 |
| $\tau_{(5–8).3}$ | 0.643/medium | −0.036 | −0.056 | 0.099 | 0.092 | −0.004 | 0.003 | 0.930 |
| $\tau_{(5–8).4}$ | 1.645/medium | 0.006 | 0.003 | 0.175 | 0.175 | −0.015 | 0.028 | 0.955 |
| $\tau_{(9–12).1}$ | −1.645/weak | −0.007 | 0.004 | 0.188 | 0.188 | −0.034 | 0.032 | 0.946 |
| $\tau_{(9–12).2}$ | −0.643/weak | 0.052 | −0.081 | 0.105 | 0.091 | −0.008 | 0.004 | 0.901 |
| $\tau_{(9–12).3}$ | 0.643/weak | −0.052 | −0.080 | 0.104 | 0.091 | −0.005 | 0.004 | 0.896 |
| $\tau_{(9–12).4}$ | 1.645/weak | 0.005 | 0.003 | 0.191 | 0.191 | −0.052 | 0.033 | 0.946 |
| $\alpha_{(1–4)}$ | 2.269/strong | 0.011 | 0.005 | 0.306 | 0.305 | 0.013 | 0.051 | 0.949 |
| $\alpha_{(5–8)}$ | 0.983/medium | −0.022 | −0.023 | 0.171 | 0.170 | −0.010 | 0.017 | 0.942 |
| $\alpha_{(9–12)}$ | 0.535/weak | −0.021 | −0.040 | 0.143 | 0.142 | 0.020 | 0.014 | 0.946 |
| $\beta_{(1–4).1}$ | −2.056/strong | −0.029 | 0.014 | 0.249 | 0.247 | | | |
| $\beta_{(1–4).2}$ | −0.804/strong | 0.010 | −0.012 | 0.130 | 0.129 | | | |
| $\beta_{(1–4).3}$ | 0.804/strong | −0.004 | −0.005 | 0.128 | 0.128 | | | |
| $\beta_{(1–4).4}$ | 2.056/strong | 0.039 | 0.019 | 0.253 | 0.250 | | | |
| $\beta_{(5–8).1}$ | −3.290/medium | −0.166 | 0.050 | 0.690 | 0.670 | | | |
| $\beta_{(5–8).2}$ | −1.287/medium | 0.019 | −0.015 | 0.269 | 0.269 | | | |
| $\beta_{(5–8).3}$ | 1.287/medium | −0.016 | −0.013 | 0.269 | 0.268 | | | |
| $\beta_{(5–8).4}$ | 3.290/medium | 0.171 | 0.052 | 0.689 | 0.667 | | | |
| $\beta_{(9–12).1}$ | −5.483/weak | −0.803 | 0.146 | 2.607 | 2.481 | | | |
| $\beta_{(9–12).2}$ | −2.144/weak | −0.095 | 0.045 | 0.886 | 0.881 | | | |
| $\beta_{(9–12).3}$ | 2.144/weak | 0.096 | 0.045 | 0.893 | 0.889 | | | |
| $\beta_{(9–12).4}$ | 5.483/weak | 0.799 | 0.146 | 2.653 | 2.531 | | | |

*Table D.11.* FA-lin standardized parameters of interest and corresponding standard errors in Cell nNM6 averaged over items with equal population value/loading. $n = 600$; $R = 1000$.

| | $\omega$ | PB($\hat{\omega}$) | RB($\hat{\omega}$) | RMSE($\hat{\omega}$) | SD($\hat{\omega}$) | RB($\hat{se}$) | RMSE($\hat{se}$) | 95%-cov. |
|---|---|---|---|---|---|---|---|---|
| $\lambda_{(1–4)}$ | 0.800/strong | −0.044 | −0.055 | 0.049 | 0.022 | −0.013 | 0.001 | 0.467 |
| $\lambda_{(5–8)}$ | 0.500/medium | −0.029 | −0.057 | 0.044 | 0.034 | 0.024 | 0.002 | 0.892 |
| $\lambda_{(9–12)}$ | 0.300/weak | −0.017 | −0.058 | 0.044 | 0.040 | −0.001 | 0.001 | 0.929 |

*Table D.12.* FA-poly standardized parameters of interest and corresponding standard errors in Cell nNM6 averaged over items with equal population value/loading. $n = 600$; $R = 1000$.

| | $\omega$ | PB($\hat{\omega}$) | RB($\hat{\omega}$) | RMSE($\hat{\omega}$) | SD($\hat{\omega}$) | RB($\hat{se}$) | RMSE($\hat{se}$) | 95%-cov. |
|---|---|---|---|---|---|---|---|---|
| $\lambda_{(1–4)}$ | 0.800/strong | 0.001 | 0.001 | 0.022 | 0.022 | −0.025 | 0.002 | 0.941 |
| $\lambda_{(5–8)}$ | 0.500/medium | 0.001 | 0.002 | 0.036 | 0.036 | 0.001 | 0.002 | 0.950 |
| $\lambda_{(9–12)}$ | 0.300/weak | 0.001 | 0.002 | 0.043 | 0.043 | −0.025 | 0.002 | 0.946 |
| $\tau_{(1–4).1}$ | −1.645/strong | −0.007 | 0.004 | 0.088 | 0.088 | −0.009 | 0.005 | 0.951 |
| $\tau_{(1–4).2}$ | −0.643/strong | 0.000 | −0.001 | 0.055 | 0.055 | −0.000 | 0.001 | 0.950 |
| $\tau_{(1–4).3}$ | 0.643/strong | 0.000 | 0.000 | 0.056 | 0.056 | −0.019 | 0.001 | 0.948 |
| $\tau_{(1–4).4}$ | 1.645/strong | 0.009 | 0.005 | 0.090 | 0.089 | −0.025 | 0.006 | 0.952 |
| $\tau_{(5–8).1}$ | −1.645/medium | −0.005 | 0.003 | 0.087 | 0.087 | −0.005 | 0.006 | 0.954 |
| $\tau_{(5–8).2}$ | −0.643/medium | 0.000 | −0.001 | 0.056 | 0.056 | −0.017 | 0.001 | 0.947 |
| $\tau_{(5–8).3}$ | 0.643/medium | 0.002 | 0.003 | 0.056 | 0.056 | −0.016 | 0.001 | 0.948 |
| $\tau_{(5–8).4}$ | 1.645/medium | 0.006 | 0.004 | 0.088 | 0.087 | −0.005 | 0.005 | 0.949 |
| $\tau_{(9–12).1}$ | −1.645/weak | −0.006 | 0.004 | 0.089 | 0.089 | −0.022 | 0.006 | 0.950 |
| $\tau_{(9–12).2}$ | −0.643/weak | −0.001 | 0.002 | 0.054 | 0.054 | 0.025 | 0.002 | 0.956 |
| $\tau_{(9–12).3}$ | 0.643/weak | 0.002 | 0.004 | 0.055 | 0.055 | −0.000 | 0.001 | 0.951 |
| $\tau_{(9–12).4}$ | 1.645/weak | 0.008 | 0.005 | 0.089 | 0.089 | −0.017 | 0.006 | 0.954 |
| $\alpha_{(1–4)}$ | 2.269/strong | 0.019 | 0.008 | 0.175 | 0.174 | −0.029 | 0.016 | 0.949 |
| $\alpha_{(5–8)}$ | 0.983/medium | 0.005 | 0.005 | 0.094 | 0.094 | 0.001 | 0.004 | 0.953 |
| $\alpha_{(9–12)}$ | 0.535/weak | 0.003 | 0.006 | 0.085 | 0.085 | −0.026 | 0.003 | 0.946 |
| $\beta_{(1–4).1}$ | −2.056/strong | −0.009 | 0.004 | 0.134 | 0.134 | | | |
| $\beta_{(1–4).2}$ | −0.804/strong | 0.001 | −0.001 | 0.075 | 0.075 | | | |
| $\beta_{(1–4).3}$ | 0.804/strong | 0.000 | 0.000 | 0.076 | 0.076 | | | |
| $\beta_{(1–4).4}$ | 2.056/strong | 0.011 | 0.005 | 0.137 | 0.136 | | | |
| $\beta_{(5–8).1}$ | −3.290/medium | −0.025 | 0.008 | 0.322 | 0.321 | | | |
| $\beta_{(5–8).2}$ | −1.287/medium | −0.005 | 0.004 | 0.154 | 0.154 | | | |
| $\beta_{(5–8).3}$ | 1.287/medium | 0.009 | 0.007 | 0.152 | 0.152 | | | |
| $\beta_{(5–8).4}$ | 3.290/medium | 0.027 | 0.008 | 0.321 | 0.320 | | | |
| $\beta_{(9–12).1}$ | −5.483/weak | −0.133 | 0.024 | 0.948 | 0.939 | | | |
| $\beta_{(9–12).2}$ | −2.144/weak | −0.048 | 0.022 | 0.395 | 0.392 | | | |
| $\beta_{(9–12).3}$ | 2.144/weak | 0.052 | 0.024 | 0.403 | 0.400 | | | |
| $\beta_{(9–12).4}$ | 5.483/weak | 0.141 | 0.026 | 0.953 | 0.943 | | | |

*Table D.13.* IRT-grm standardized parameters of interest and corresponding standard errors in Cell nNM6 averaged over items with equal population value/loading. $n = 600$; $R = 1000$.

| | $\omega$ | PB($\hat{\omega}$) | RB($\hat{\omega}$) | RMSE($\hat{\omega}$) | SD($\hat{\omega}$) | RB($\hat{se}$) | RMSE($\hat{se}$) | 95%-cov. |
|---|---|---|---|---|---|---|---|---|
| $\lambda_{(1–4)}$ | 0.800/strong | −0.003 | −0.004 | 0.023 | 0.022 | −0.011 | 0.002 | 0.951 |
| $\lambda_{(5–8)}$ | 0.500/medium | −0.013 | −0.026 | 0.039 | 0.036 | 0.023 | 0.002 | 0.946 |
| $\lambda_{(9–12)}$ | 0.300/weak | −0.012 | −0.041 | 0.044 | 0.042 | 0.007 | 0.002 | 0.942 |
| $\tau_{(1–4).1}$ | −1.645/strong | −0.001 | 0.000 | 0.089 | 0.089 | −0.007 | 0.006 | 0.948 |
| $\tau_{(1–4).2}$ | −0.643/strong | 0.012 | −0.018 | 0.055 | 0.054 | −0.003 | 0.001 | 0.940 |
| $\tau_{(1–4).3}$ | 0.643/strong | −0.011 | −0.017 | 0.056 | 0.055 | −0.015 | 0.002 | 0.943 |
| $\tau_{(1–4).4}$ | 1.645/strong | 0.002 | 0.001 | 0.090 | 0.090 | −0.022 | 0.007 | 0.946 |
| $\tau_{(5–8).1}$ | −1.645/medium | 0.014 | −0.008 | 0.098 | 0.097 | −0.006 | 0.008 | 0.946 |
| $\tau_{(5–8).2}$ | −0.643/medium | 0.041 | −0.064 | 0.067 | 0.053 | −0.015 | 0.001 | 0.866 |
| $\tau_{(5–8).3}$ | 0.643/medium | −0.039 | −0.061 | 0.066 | 0.053 | −0.016 | 0.001 | 0.875 |
| $\tau_{(5–8).4}$ | 1.645/medium | −0.013 | −0.008 | 0.098 | 0.098 | −0.006 | 0.008 | 0.942 |
| $\tau_{(9–12).1}$ | −1.645/weak | 0.014 | −0.008 | 0.105 | 0.104 | −0.024 | 0.010 | 0.940 |
| $\tau_{(9–12).2}$ | −0.643/weak | 0.055 | −0.086 | 0.075 | 0.051 | 0.022 | 0.002 | 0.811 |
| $\tau_{(9–12).3}$ | 0.643/weak | −0.054 | −0.084 | 0.075 | 0.052 | −0.000 | 0.001 | 0.809 |
| $\tau_{(9–12).4}$ | 1.645/weak | −0.011 | −0.007 | 0.105 | 0.104 | −0.019 | 0.009 | 0.940 |
| $\alpha_{(1–4)}$ | 2.269/strong | −0.013 | −0.006 | 0.176 | 0.176 | −0.015 | 0.016 | 0.943 |
| $\alpha_{(5–8)}$ | 0.983/medium | −0.031 | −0.031 | 0.098 | 0.093 | 0.022 | 0.006 | 0.938 |
| $\alpha_{(9–12)}$ | 0.535/weak | −0.022 | −0.041 | 0.085 | 0.082 | 0.005 | 0.004 | 0.939 |
| $\beta_{(1–4).1}$ | −2.056/strong | −0.011 | 0.006 | 0.137 | 0.136 | | | |
| $\beta_{(1–4).2}$ | −0.804/strong | 0.011 | −0.013 | 0.074 | 0.073 | | | |
| $\beta_{(1–4).3}$ | 0.804/strong | −0.010 | −0.012 | 0.075 | 0.074 | | | |
| $\beta_{(1–4).4}$ | 2.056/strong | 0.013 | 0.006 | 0.138 | 0.138 | | | |
| $\beta_{(5–8).1}$ | −3.290/medium | −0.081 | 0.025 | 0.359 | 0.350 | | | |
| $\beta_{(5–8).2}$ | −1.287/medium | 0.043 | −0.033 | 0.153 | 0.147 | | | |
| $\beta_{(5–8).3}$ | 1.287/medium | −0.039 | −0.030 | 0.150 | 0.145 | | | |
| $\beta_{(5–8).4}$ | 3.290/medium | 0.083 | 0.025 | 0.358 | 0.349 | | | |
| $\beta_{(9–12).1}$ | −5.483/weak | −0.317 | 0.058 | 1.060 | 1.011 | | | |
| $\beta_{(9–12).2}$ | −2.144/weak | 0.055 | −0.025 | 0.377 | 0.373 | | | |
| $\beta_{(9–12).3}$ | 2.144/weak | −0.051 | −0.024 | 0.383 | 0.379 | | | |
| $\beta_{(9–12).4}$ | 5.483/weak | 0.328 | 0.060 | 1.069 | 1.018 | | | |

### D.4.2 Coverage Results for $\lambda$ and $\tau$

Coverage rates of loading $\lambda$ and threshold $\tau$ parameter estimators for all parametric models in all basic data configurations are given in Table D.14. These coverage rates are averaged over parameters that belong to items that have equal loadings and equal population values. As a result, the coverage rates are based on various numbers of replications. For example, whereas the average coverage rate of $\lambda_{1–12}$ in Cell nNS2 is based on $R = 12000$ replications, the average coverage rate of $\lambda_{1–4}$ in Cell nNM2 is based on $R = 4000$ replications. The numbers are *deviations* from the expected coverage rate, i.e., 0.95 is subtracted from the empirical average. Coverage rates are considered acceptable when they are between 0.90 and 0.98, corresponding to a deviation between −0.05 and 0.03. Coverage rates beyond these limits are printed in boldface.

*Table D.14.* Coverage rate of 95%-confidence interval estimators for Cells nNS2, nNS6, nNM2, and nNM6 averaged over parameters with equal population value/loading. $R = 1000$ per parameter. Numbers represent deviation from 0.95, with values outside the range $(-0.05, 0.03)$ printed in boldface, indicating an unacceptable coverage rate.

| Cell | | $\omega$ | FA-lin | FA-poly | IRT-grm |
|---|---|---|---|---|---|
| nNS2 | $\lambda_{(1–12)}$ | 0.800/strong | **−0.170** | −0.023 | −0.004 |
| | $\tau_{(1–12).1}$ | −1.645/strong | | −0.001 | 0.002 |
| | $\tau_{(1–12).2}$ | −0.643/strong | | −0.002 | −0.002 |
| | $\tau_{(1–12).3}$ | 0.643/strong | | −0.003 | −0.005 |
| | $\tau_{(1–12).4}$ | 1.645/strong | | −0.003 | −0.002 |
| nNS6 | $\lambda_{(1–12)}$ | 0.800/strong | **−0.597** | −0.007 | 0.002 |
| | $\tau_{(1–12).1}$ | −1.645/strong | | 0.001 | −0.002 |
| | $\tau_{(1–12).2}$ | −0.643/strong | | −0.006 | −0.009 |
| | $\tau_{(1–12).3}$ | 0.643/strong | | −0.004 | −0.005 |
| | $\tau_{(1–12).4}$ | 1.645/strong | | 0.000 | −0.003 |
| nNM2 | $\lambda_{(1–4)}$ | 0.800/strong | **−0.120** | −0.014 | −0.002 |
| | $\lambda_{(5–8)}$ | 0.500/medium | −0.014 | −0.028 | −0.005 |
| | $\lambda_{(9–12)}$ | 0.300/weak | −0.009 | −0.023 | −0.006 |
| | $\tau_{(1–4).1}$ | −1.645/strong | | −0.000 | −0.001 |
| | $\tau_{(1–4).2}$ | −0.643/strong | | −0.004 | −0.008 |
| | $\tau_{(1–4).3}$ | 0.643/strong | | −0.005 | −0.005 |
| | $\tau_{(1–4).4}$ | 1.645/strong | | 0.000 | 0.001 |
| | $\tau_{(5–8).1}$ | −1.645/medium | | 0.003 | −0.001 |
| | $\tau_{(5–8).2}$ | −0.643/medium | | −0.008 | −0.030 |
| | $\tau_{(5–8).3}$ | 0.643/medium | | −0.002 | −0.020 |
| | $\tau_{(5–8).4}$ | 1.645/medium | | 0.002 | 0.005 |
| | $\tau_{(9–12).1}$ | −1.645/weak | | −0.002 | −0.004 |
| | $\tau_{(9–12).2}$ | −0.643/weak | | −0.003 | −0.048 |
| | $\tau_{(9–12).3}$ | 0.643/weak | | 0.001 | **−0.054** |
| | $\tau_{(9–12).4}$ | 1.645/weak | | −0.004 | −0.004 |
| nNM6 | $\lambda_{(1–4)}$ | 0.800/strong | **−0.483** | −0.009 | 0.001 |
| | $\lambda_{(5–8)}$ | 0.500/medium | **−0.058** | 0.000 | −0.004 |
| | $\lambda_{(9–12)}$ | 0.300/weak | −0.021 | −0.004 | −0.008 |
| | $\tau_{(1–4).1}$ | −1.645/strong | | 0.001 | −0.002 |
| | $\tau_{(1–4).2}$ | −0.643/strong | | 0.000 | −0.010 |
| | $\tau_{(1–4).3}$ | 0.643/strong | | −0.002 | −0.007 |
| | $\tau_{(1–4).4}$ | 1.645/strong | | 0.002 | −0.004 |
| | $\tau_{(5–8).1}$ | −1.645/medium | | 0.004 | −0.004 |
| | $\tau_{(5–8).2}$ | −0.643/medium | | −0.003 | **−0.084** |
| | $\tau_{(5–8).3}$ | 0.643/medium | | −0.002 | **−0.075** |
| | $\tau_{(5–8).4}$ | 1.645/medium | | −0.001 | −0.008 |
| | $\tau_{(9–12).1}$ | −1.645/weak | | 0.000 | −0.010 |
| | $\tau_{(9–12).2}$ | −0.643/weak | | 0.006 | **−0.139** |
| | $\tau_{(9–12).3}$ | 0.643/weak | | 0.001 | **−0.140** |
| | $\tau_{(9–12).4}$ | 1.645/weak | | 0.004 | −0.010 |

### D.4.3 Average Loevinger's $H$ Results for IRT-mok

*Table D.15.* IRT-mok $H_i$ results for Cell nNS2 averaged over parameters with equal population value. $n = 200$; $R = 1000$.

| | $\omega$ | $\bar{\hat{\omega}}$ | PB($\hat{\omega}$) | RB($\hat{\omega}$) | RMSE($\hat{\omega}$) | SD($\hat{\omega}$) | RB($\hat{se}$) | RMSE($\hat{se}$) | 95%-cov. |
|---|---|---|---|---|---|---|---|---|---|
| $H_{(1–12)}$ | 0.571 | 0.599 | 0.027 | 0.048 | 0.043 | 0.033 | 0.027 | 0.003 | 0.859 |
| $H_{scale}$ | 0.571 | 0.599 | 0.027 | 0.048 | 0.038 | 0.026 | 0.030 | 0.002 | 0.827 |

*Table D.16.* IRT-mok $H_i$ results for Cell nNS6 averaged over parameters with equal population value. $n = 600$; $R = 1000$.

| | $\omega$ | $\bar{\hat{\omega}}$ | PB($\hat{\omega}$) | RB($\hat{\omega}$) | RMSE($\hat{\omega}$) | SD($\hat{\omega}$) | RB($\hat{se}$) | RMSE($\hat{se}$) | 95%-cov. |
|---|---|---|---|---|---|---|---|---|---|
| $H_{(1–12)}$ | 0.571 | 0.588 | 0.017 | 0.029 | 0.026 | 0.019 | 0.021 | 0.001 | 0.856 |
| $H_{scale}$ | 0.571 | 0.588 | 0.017 | 0.029 | 0.023 | 0.016 | 0.006 | 0.001 | 0.799 |

*Table D.17.* IRT-mok $H_i$ results for Cell nNM2 averaged over parameters with equal population value. $n = 200$; $R = 1000$.

| | $\omega$ | $\bar{\hat{\omega}}$ | PB($\hat{\omega}$) | RB($\hat{\omega}$) | RMSE($\hat{\omega}$) | SD($\hat{\omega}$) | RB($\hat{se}$) | RMSE($\hat{se}$) | 95%-cov. |
|---|---|---|---|---|---|---|---|---|---|
| $H_{(1–4)}$ | 0.363 | 0.382 | 0.019 | 0.051 | 0.036 | 0.031 | −0.005 | 0.002 | 0.905 |
| $H_{(5–8)}$ | 0.239 | 0.252 | 0.013 | 0.053 | 0.041 | 0.039 | −0.027 | 0.003 | 0.928 |
| $H_{(9–12)}$ | 0.148 | 0.156 | 0.008 | 0.051 | 0.042 | 0.042 | −0.007 | 0.003 | 0.938 |
| $H_{scale}$ | 0.250 | 0.264 | 0.013 | 0.053 | 0.029 | 0.026 | −0.040 | 0.002 | 0.911 |

*Table D.18.* IRT-mok $H_i$ results for Cell nNM6 averaged over parameters with equal population value. $n = 600$; $R = 1000$.

| | $\omega$ | $\bar{\hat{\omega}}$ | PB($\hat{\omega}$) | RB($\hat{\omega}$) | RMSE($\hat{\omega}$) | SD($\hat{\omega}$) | RB($\hat{se}$) | RMSE($\hat{se}$) | 95%-cov. |
|---|---|---|---|---|---|---|---|---|---|
| $H_{(1–4)}$ | 0.363 | 0.375 | 0.012 | 0.032 | 0.021 | 0.017 | 0.012 | 0.001 | 0.897 |
| $H_{(5–8)}$ | 0.239 | 0.247 | 0.008 | 0.033 | 0.023 | 0.021 | 0.017 | 0.001 | 0.938 |
| $H_{(9–12)}$ | 0.148 | 0.153 | 0.005 | 0.035 | 0.024 | 0.024 | −0.007 | 0.001 | 0.940 |
| $H_{scale}$ | 0.250 | 0.259 | 0.008 | 0.033 | 0.016 | 0.014 | −0.004 | 0.001 | 0.908 |

## D.5 Precision of Reported Estimates

In choosing the number of digits to report for the various point estimates, we loosely applied the leading-digit rule with parameter $a = 1$, i.e., LDR(1); see Song and Schmeiser (2008, 2009). Doing so ensures that each nonreported digit has a probability of being correct smaller than 0.117 which is just above the baseline random probability of 0.1, since there are ten possible digits (0–9). The LDR(1) can be applied as follows. The standard error of a point estimate $\hat{\omega}$ is computed by dividing its standard deviation by the square root of the number of replications,

$$\hat{se}(\hat{\omega}) = \frac{sd(\hat{\omega})}{\sqrt{R}}. \tag{D.1}$$

The decimal place of the leading digit, i.e., the first nonzero digit, of this number is taken as the last meaningful decimal place in the point estimate.

We did not apply the LDR(1) in the strict sense, i.e., by computing it for each point estimate separately. Rather, since we prefer a consistent display of results, we determined a reasonable number of decimal places for all point estimates. To this end, we computed the boundaries in standard deviation values leading to significance of an additional decimal place. This information is provided in Table D.19 for five exemplary boundary values presented in the third column. For example, to justify reporting of four decimal places of $\hat{\omega}$, its standard error should be at most 0.0001. By applying Equation D.1, we obtain that $sd(\hat{\omega})$ is at least 0.00316 then. Since in our simulation results most point estimate standard deviations $sd(\hat{\omega})$ are between 0.0316 and 0.316, all point estimates $\hat{\omega}$ are reported up to the third decimal place.

*Table D.19.* Number of decimal places to report of point estimates $\hat{\omega}$ for various values of its standard deviation $sd(\hat{\omega})$ when $R = 1000$.

| # Decimal places | Standard error of point estimate | Standard deviation of point estimate |
|---|---|---|
| 0 | 1.0 | $\geq$ 31.62278 |
| 1 | 0.1 | $\geq$ 3.16228 |
| 2 | 0.01 | $\geq$ 0.31623 |
| 3 | 0.001 | $\geq$ 0.03162 |
| 4 | 0.0001 | $\geq$ 0.00316 |

## D.6 Additional Fit Results: $\chi^2_{YB}$

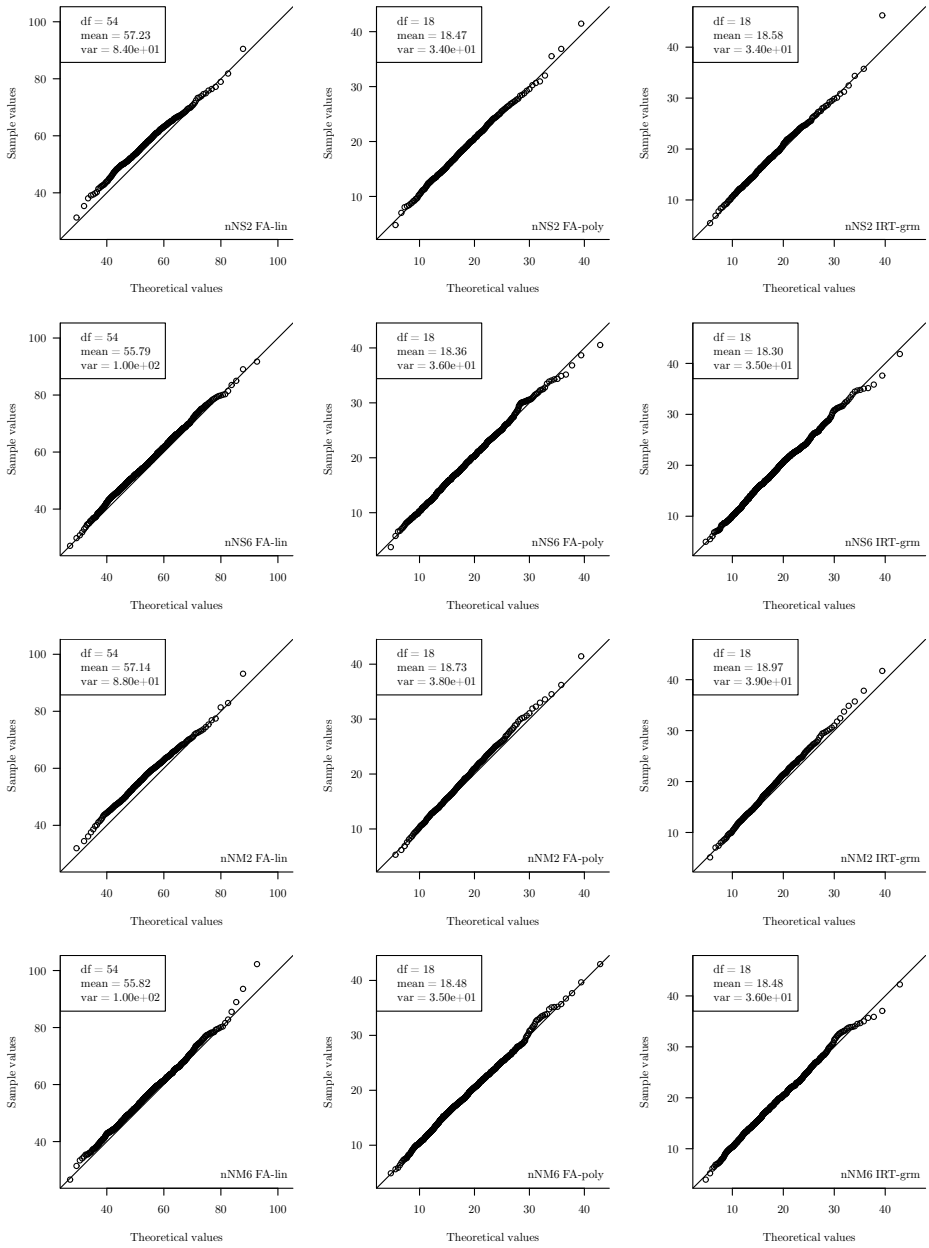



*Figure D.3.* Q-Q plots for $\chi^2_{YB}$ fit statistic for Cells nNS2, nNS6, nNM2, and nNM6 and each model. $R = 1000$. The diagonal line depicts a perfect association between the empirical and theoretical distributions.

# Appendix E

# Simulation Study: Violations of Assumptions

## E.1 Tables of Parameter and Standard Error Estimates

### E.1.1 Average Parameter and Standard Error Results for all Parameters

Parameter estimates are averaged over parameters with equal loading values (strong, medium, or weak). The coverage rate of the parameter's 95%-confidence interval is presented under 95%-cov.

*Table E.1.* FA-lin standardized parameters of interest and corresponding standard errors in Cell nNS2 averaged over items with equal population value/shape. $n = 200$; $R = 1000$.

| | $\omega$ | PB($\hat{\omega}$) | RB($\hat{\omega}$) | RMSE($\hat{\omega}$) | SD($\hat{\omega}$) | RB($\hat{se}$) | RMSE($\hat{se}$) | 95%-cov. |
|---|---|---|---|---|---|---|---|---|
| $\lambda_{(1–12)}$ | 0.800/normal | −0.045 | −0.056 | 0.056 | 0.033 | −0.002 | 0.004 | 0.780 |

*Table E.2.* FA-poly standardized parameters of interest and corresponding standard errors in Cell nNS2 averaged over items with equal population value/shape. $n = 200$; $R = 1000$.

| | $\omega$ | PB($\hat{\omega}$) | RB($\hat{\omega}$) | RMSE($\hat{\omega}$) | SD($\hat{\omega}$) | RB($\hat{se}$) | RMSE($\hat{se}$) | 95%-cov. |
|---|---|---|---|---|---|---|---|---|
| $\lambda_{(1–12)}$ | 0.800/normal | 0.003 | 0.003 | 0.033 | 0.032 | −0.052 | 0.004 | 0.927 |
| $\tau_{(1–12).1}$ | −1.645/normal | −0.023 | 0.014 | 0.163 | 0.161 | −0.044 | 0.020 | 0.949 |
| $\tau_{(1–12).2}$ | −0.643/normal | −0.002 | 0.004 | 0.096 | 0.096 | −0.001 | 0.003 | 0.948 |
| $\tau_{(1–12).3}$ | 0.643/normal | 0.004 | 0.006 | 0.096 | 0.096 | −0.003 | 0.003 | 0.947 |
| $\tau_{(1–12).4}$ | 1.645/normal | 0.020 | 0.012 | 0.161 | 0.160 | −0.040 | 0.020 | 0.947 |
| $\alpha_{(1–12)}$ | 2.269/normal | 0.049 | 0.022 | 0.270 | 0.266 | −0.054 | 0.033 | 0.944 |
| $\beta_{(1–12).1}$ | −2.056/normal | −0.028 | 0.013 | 0.239 | 0.237 | | | |
| $\beta_{(1–12).2}$ | −0.804/normal | −0.002 | 0.003 | 0.129 | 0.129 | | | |
| $\beta_{(1–12).3}$ | 0.804/normal | 0.004 | 0.005 | 0.129 | 0.129 | | | |
| $\beta_{(1–12).4}$ | 2.056/normal | 0.024 | 0.012 | 0.238 | 0.237 | | | |

*Table E.3.* IRT-grm standardized parameters of interest and corresponding standard errors in Cell nNS2 averaged over items with equal population value/shape. $n = 200$; $R = 1000$.

| | $\omega$ | PB($\hat{\omega}$) | RB($\hat{\omega}$) | RMSE($\hat{\omega}$) | SD($\hat{\omega}$) | RB($\hat{se}$) | RMSE($\hat{se}$) | 95%-cov. |
|---|---|---|---|---|---|---|---|---|
| $\lambda_{(1–12)}$ | 0.800/normal | −0.003 | −0.004 | 0.034 | 0.034 | −0.010 | 0.004 | 0.946 |
| $\tau_{(1–12).1}$ | −1.645/normal | −0.016 | 0.010 | 0.165 | 0.164 | −0.054 | 0.024 | 0.952 |
| $\tau_{(1–12).2}$ | −0.643/normal | 0.009 | −0.014 | 0.094 | 0.094 | −0.006 | 0.004 | 0.948 |
| $\tau_{(1–12).3}$ | 0.643/normal | −0.008 | −0.012 | 0.095 | 0.094 | −0.012 | 0.004 | 0.945 |
| $\tau_{(1–12).4}$ | 1.645/normal | 0.014 | 0.008 | 0.165 | 0.164 | −0.056 | 0.024 | 0.948 |
| $\alpha_{(1–12)}$ | 2.269/normal | 0.004 | 0.002 | 0.268 | 0.268 | −0.012 | 0.033 | 0.946 |
| $\beta_{(1–12).1}$ | −2.056/normal | −0.034 | 0.017 | 0.246 | 0.244 | | | |
| $\beta_{(1–12).2}$ | −0.804/normal | 0.006 | −0.007 | 0.126 | 0.126 | | | |
| $\beta_{(1–12).3}$ | 0.804/normal | −0.005 | −0.006 | 0.126 | 0.126 | | | |
| $\beta_{(1–12).4}$ | 2.056/normal | 0.031 | 0.015 | 0.246 | 0.244 | | | |

*Table E.4.* FA-lin standardized parameters of interest and corresponding standard errors in Cell nNS6 averaged over items with equal population value/shape. $n = 600$; $R = 1000$.

| | $\omega$ | PB($\hat{\omega}$) | RB($\hat{\omega}$) | RMSE($\hat{\omega}$) | SD($\hat{\omega}$) | RB($\hat{se}$) | RMSE($\hat{se}$) | 95%-cov. |
|---|---|---|---|---|---|---|---|---|
| $\lambda_{(1–12)}$ | 0.800/normal | −0.044 | −0.055 | 0.048 | 0.019 | 0.003 | 0.001 | 0.353 |

*Table E.5.* FA-poly standardized parameters of interest and corresponding standard errors in Cell nNS6 averaged over items with equal population value/shape. $n = 600$; $R = 1000$.

| | $\omega$ | PB($\hat{\omega}$) | RB($\hat{\omega}$) | RMSE($\hat{\omega}$) | SD($\hat{\omega}$) | RB($\hat{se}$) | RMSE($\hat{se}$) | 95%-cov. |
|---|---|---|---|---|---|---|---|---|
| $\lambda_{(1–12)}$ | 0.800/normal | 0.001 | 0.002 | 0.019 | 0.019 | −0.013 | 0.001 | 0.943 |
| $\tau_{(1–12).1}$ | −1.645/normal | −0.004 | 0.003 | 0.088 | 0.087 | −0.007 | 0.005 | 0.951 |
| $\tau_{(1–12).2}$ | −0.643/normal | −0.000 | 0.000 | 0.056 | 0.056 | −0.014 | 0.001 | 0.944 |
| $\tau_{(1–12).3}$ | 0.643/normal | 0.000 | 0.001 | 0.056 | 0.056 | −0.014 | 0.002 | 0.946 |
| $\tau_{(1–12).4}$ | 1.645/normal | 0.004 | 0.003 | 0.088 | 0.087 | −0.006 | 0.006 | 0.950 |
| $\alpha_{(1–12)}$ | 2.269/normal | 0.019 | 0.008 | 0.149 | 0.148 | −0.009 | 0.010 | 0.949 |
| $\beta_{(1–12).1}$ | −2.056/normal | −0.004 | 0.002 | 0.131 | 0.131 | | | |
| $\beta_{(1–12).2}$ | −0.804/normal | 0.000 | −0.000 | 0.075 | 0.075 | | | |
| $\beta_{(1–12).3}$ | 0.804/normal | −0.000 | −0.000 | 0.075 | 0.075 | | | |
| $\beta_{(1–12).4}$ | 2.056/normal | 0.004 | 0.002 | 0.130 | 0.130 | | | |

*Table E.6.* IRT-grm standardized parameters of interest and corresponding standard errors in Cell nNS6 averaged over items with equal population value/shape. $n = 600$; $R = 1000$.

| | $\omega$ | PB($\hat{\omega}$) | RB($\hat{\omega}$) | RMSE($\hat{\omega}$) | SD($\hat{\omega}$) | RB($\hat{se}$) | RMSE($\hat{se}$) | 95%-cov. |
|---|---|---|---|---|---|---|---|---|
| $\lambda_{(1–12)}$ | 0.800/normal | −0.004 | −0.005 | 0.020 | 0.019 | 0.003 | 0.002 | 0.952 |
| $\tau_{(1–12).1}$ | −1.645/normal | 0.003 | −0.002 | 0.088 | 0.088 | −0.008 | 0.006 | 0.948 |
| $\tau_{(1–12).2}$ | −0.643/normal | 0.010 | −0.016 | 0.055 | 0.054 | −0.008 | 0.001 | 0.941 |
| $\tau_{(1–12).3}$ | 0.643/normal | −0.010 | −0.016 | 0.055 | 0.054 | −0.005 | 0.002 | 0.945 |
| $\tau_{(1–12).4}$ | 1.645/normal | −0.002 | −0.001 | 0.088 | 0.088 | −0.009 | 0.006 | 0.947 |
| $\alpha_{(1–12)}$ | 2.269/normal | −0.020 | −0.009 | 0.150 | 0.149 | 0.007 | 0.011 | 0.944 |
| $\beta_{(1–12).1}$ | −2.056/normal | −0.008 | 0.004 | 0.133 | 0.132 | | | |
| $\beta_{(1–12).2}$ | −0.804/normal | 0.009 | −0.011 | 0.073 | 0.073 | | | |
| $\beta_{(1–12).3}$ | 0.804/normal | −0.009 | −0.011 | 0.073 | 0.072 | | | |
| $\beta_{(1–12).4}$ | 2.056/normal | 0.009 | 0.004 | 0.132 | 0.132 | | | |

*Table E.7.* FA-lin standardized parameters of interest and corresponding standard errors in Cell rnNS2 averaged over items with equal population value/shape. $n = 200$; $R = 1000$.

| | $\omega$ | PB($\hat{\omega}$) | RB($\hat{\omega}$) | RMSE($\hat{\omega}$) | SD($\hat{\omega}$) | RB($\hat{se}$) | RMSE($\hat{se}$) | 95%-cov. |
|---|---|---|---|---|---|---|---|---|
| $\lambda_{(1–6)}$ | 0.800/normal | −0.061 | −0.076 | 0.070 | 0.034 | 0.027 | 0.004 | 0.638 |
| $\lambda_{(7–12)}$ | 0.800/right-skewed | −0.096 | −0.120 | 0.106 | 0.044 | −0.123 | 0.007 | 0.292 |

*Table E.8.* FA-poly standardized parameters of interest and corresponding standard errors in Cell rnNS2 averaged over items with equal population value/shape. $n = 200$; $R = 1000$.

| | $\omega$ | PB($\hat{\omega}$) | RB($\hat{\omega}$) | RMSE($\hat{\omega}$) | SD($\hat{\omega}$) | RB($\hat{se}$) | RMSE($\hat{se}$) | 95%-cov. |
|---|---|---|---|---|---|---|---|---|
| $\lambda_{(1–6)}$ | 0.800/normal | 0.003 | 0.004 | 0.033 | 0.033 | −0.047 | 0.005 | 0.918 |
| $\lambda_{(7–12)}$ | 0.800/right-skewed | 0.003 | 0.004 | 0.040 | 0.040 | −0.057 | 0.006 | 0.918 |
| $\tau_{(1–6).1}$ | −1.645/normal | −0.018 | 0.011 | 0.161 | 0.160 | −0.045 | 0.020 | 0.949 |
| $\tau_{(1–6).2}$ | −0.643/normal | −0.003 | 0.004 | 0.097 | 0.097 | −0.013 | 0.003 | 0.946 |
| $\tau_{(1–6).3}$ | 0.643/normal | 0.002 | 0.003 | 0.097 | 0.097 | −0.012 | 0.003 | 0.944 |
| $\tau_{(1–6).4}$ | 1.645/normal | 0.017 | 0.010 | 0.157 | 0.157 | −0.023 | 0.018 | 0.954 |
| $\tau_{(7–12).1}$ | 0.346/right-skewed | 0.001 | 0.003 | 0.094 | 0.094 | −0.030 | 0.003 | 0.946 |
| $\tau_{(7–12).2}$ | 0.800/right-skewed | 0.002 | 0.002 | 0.103 | 0.103 | −0.027 | 0.004 | 0.938 |
| $\tau_{(7–12).3}$ | 1.221/right-skewed | 0.005 | 0.004 | 0.120 | 0.120 | −0.016 | 0.007 | 0.953 |
| $\tau_{(7–12).4}$ | 1.622/right-skewed | 0.014 | 0.008 | 0.157 | 0.157 | −0.041 | 0.019 | 0.943 |
| $\alpha_{(1–6)}$ | 2.269/normal | 0.058 | 0.026 | 0.280 | 0.274 | −0.049 | 0.035 | 0.940 |
| $\alpha_{(7–12)}$ | 2.269/right-skewed | 0.069 | 0.031 | 0.345 | 0.338 | −0.069 | 0.060 | 0.946 |
| $\beta_{(1–6).1}$ | −2.056/normal | −0.019 | 0.009 | 0.235 | 0.235 | | | |
| $\beta_{(1–6).2}$ | −0.804/normal | −0.002 | 0.003 | 0.130 | 0.130 | | | |
| $\beta_{(1–6).3}$ | 0.804/normal | 0.001 | 0.001 | 0.130 | 0.130 | | | |
| $\beta_{(1–6).4}$ | 2.056/normal | 0.018 | 0.009 | 0.236 | 0.235 | | | |
| $\beta_{(7–12).1}$ | 0.432/right-skewed | 0.001 | 0.001 | 0.119 | 0.119 | | | |
| $\beta_{(7–12).2}$ | 1.000/right-skewed | 0.002 | 0.002 | 0.146 | 0.146 | | | |
| $\beta_{(7–12).3}$ | 1.527/right-skewed | 0.007 | 0.005 | 0.190 | 0.190 | | | |
| $\beta_{(7–12).4}$ | 2.027/right-skewed | 0.018 | 0.009 | 0.253 | 0.253 | | | |

*Table E.9.* IRT-grm standardized parameters of interest and corresponding standard errors in Cell rnNS2 averaged over items with equal population value/shape. $n = 200$; $R = 1000$.

| | $\omega$ | PB($\hat{\omega}$) | RB($\hat{\omega}$) | RMSE($\hat{\omega}$) | SD($\hat{\omega}$) | RB($\hat{se}$) | RMSE($\hat{se}$) | 95%-cov. |
|---|---|---|---|---|---|---|---|---|
| $\lambda_{(1–6)}$ | 0.800/normal | −0.003 | −0.003 | 0.034 | 0.034 | 0.002 | 0.005 | 0.942 |
| $\lambda_{(7–12)}$ | 0.800/right-skewed | −0.007 | −0.009 | 0.041 | 0.040 | −0.002 | 0.006 | 0.944 |
| $\tau_{(1–6).1}$ | −1.645/normal | −0.015 | 0.009 | 0.163 | 0.163 | −0.046 | 0.022 | 0.954 |
| $\tau_{(1–6).2}$ | −0.643/normal | 0.008 | −0.013 | 0.095 | 0.095 | −0.014 | 0.004 | 0.946 |
| $\tau_{(1–6).3}$ | 0.643/normal | −0.010 | −0.016 | 0.094 | 0.093 | −0.004 | 0.004 | 0.945 |
| $\tau_{(1–6).4}$ | 1.645/normal | 0.007 | 0.004 | 0.159 | 0.159 | −0.030 | 0.021 | 0.953 |
| $\tau_{(7–12).1}$ | 0.346/right-skewed | −0.008 | −0.022 | 0.090 | 0.090 | −0.024 | 0.003 | 0.943 |
| $\tau_{(7–12).2}$ | 0.800/right-skewed | −0.018 | −0.022 | 0.100 | 0.099 | −0.021 | 0.005 | 0.938 |
| $\tau_{(7–12).3}$ | 1.221/right-skewed | −0.017 | −0.014 | 0.119 | 0.118 | −0.012 | 0.009 | 0.941 |
| $\tau_{(7–12).4}$ | 1.622/right-skewed | 0.000 | 0.000 | 0.158 | 0.159 | −0.044 | 0.022 | 0.947 |
| $\alpha_{(1–6)}$ | 2.269/normal | 0.010 | 0.005 | 0.271 | 0.271 | −0.001 | 0.034 | 0.949 |
| $\alpha_{(7–12)}$ | 2.269/right-skewed | −0.014 | −0.006 | 0.315 | 0.314 | −0.015 | 0.048 | 0.943 |
| $\beta_{(1–6).1}$ | −2.056/normal | −0.031 | 0.015 | 0.243 | 0.241 | | | |
| $\beta_{(1–6).2}$ | −0.804/normal | 0.006 | −0.008 | 0.127 | 0.127 | | | |
| $\beta_{(1–6).3}$ | 0.804/normal | −0.009 | −0.011 | 0.126 | 0.125 | | | |
| $\beta_{(1–6).4}$ | 2.056/normal | 0.021 | 0.010 | 0.240 | 0.239 | | | |
| $\beta_{(7–12).1}$ | 0.432/right-skewed | −0.005 | −0.011 | 0.115 | 0.115 | | | |
| $\beta_{(7–12).2}$ | 1.000/right-skewed | −0.011 | −0.011 | 0.140 | 0.139 | | | |
| $\beta_{(7–12).3}$ | 1.527/right-skewed | −0.003 | −0.002 | 0.184 | 0.184 | | | |
| $\beta_{(7–12).4}$ | 2.027/right-skewed | 0.027 | 0.013 | 0.254 | 0.253 | | | |

*Table E.10.* FA-lin standardized parameters of interest and corresponding standard errors in Cell rnNS6 averaged over items with equal population value/shape. $n = 600$; $R = 1000$.

| | $\omega$ | PB($\hat{\omega}$) | RB($\hat{\omega}$) | RMSE($\hat{\omega}$) | SD($\hat{\omega}$) | RB($\hat{se}$) | RMSE($\hat{se}$) | 95%-cov. |
|---|---|---|---|---|---|---|---|---|
| $\lambda_{(1–6)}$ | 0.800/normal | −0.061 | −0.076 | 0.064 | 0.020 | 0.017 | 0.001 | 0.113 |
| $\lambda_{(7–12)}$ | 0.800/right-skewed | −0.096 | −0.120 | 0.099 | 0.026 | −0.135 | 0.004 | 0.005 |

*Table E.11.* FA-poly standardized parameters of interest and corresponding standard errors in Cell rnNS6 averaged over items with equal population value/shape. $n = 600$; $R = 1000$.

| | $\omega$ | PB($\hat{\omega}$) | RB($\hat{\omega}$) | RMSE($\hat{\omega}$) | SD($\hat{\omega}$) | RB($\hat{se}$) | RMSE($\hat{se}$) | 95%-cov. |
|---|---|---|---|---|---|---|---|---|
| $\lambda_{(1–6)}$ | 0.800/normal | 0.001 | 0.001 | 0.019 | 0.019 | −0.027 | 0.002 | 0.939 |
| $\lambda_{(7–12)}$ | 0.800/right-skewed | 0.001 | 0.002 | 0.023 | 0.023 | −0.036 | 0.002 | 0.935 |
| $\tau_{(1–6).1}$ | −1.645/normal | −0.005 | 0.003 | 0.089 | 0.089 | −0.018 | 0.006 | 0.947 |
| $\tau_{(1–6).2}$ | −0.643/normal | −0.000 | 0.000 | 0.056 | 0.056 | −0.006 | 0.001 | 0.946 |
| $\tau_{(1–6).3}$ | 0.643/normal | 0.001 | 0.001 | 0.056 | 0.056 | −0.015 | 0.001 | 0.948 |
| $\tau_{(1–6).4}$ | 1.645/normal | 0.006 | 0.004 | 0.089 | 0.089 | −0.025 | 0.006 | 0.951 |
| $\tau_{(7–12).1}$ | 0.346/right-skewed | −0.000 | −0.000 | 0.052 | 0.052 | 0.004 | 0.001 | 0.951 |
| $\tau_{(7–12).2}$ | 0.800/right-skewed | 0.001 | 0.001 | 0.058 | 0.058 | −0.005 | 0.002 | 0.946 |
| $\tau_{(7–12).3}$ | 1.221/right-skewed | 0.004 | 0.003 | 0.069 | 0.069 | −0.013 | 0.003 | 0.945 |
| $\tau_{(7–12).4}$ | 1.622/right-skewed | 0.005 | 0.003 | 0.088 | 0.088 | −0.029 | 0.006 | 0.941 |
| $\alpha_{(1–6)}$ | 2.269/normal | 0.017 | 0.008 | 0.157 | 0.156 | −0.030 | 0.012 | 0.948 |
| $\alpha_{(7–12)}$ | 2.269/right-skewed | 0.024 | 0.011 | 0.189 | 0.188 | −0.033 | 0.018 | 0.948 |
| $\beta_{(1–6).1}$ | −2.056/normal | −0.006 | 0.003 | 0.131 | 0.131 | | | |
| $\beta_{(1–6).2}$ | −0.804/normal | 0.000 | −0.000 | 0.075 | 0.075 | | | |
| $\beta_{(1–6).3}$ | 0.804/normal | 0.001 | 0.001 | 0.076 | 0.076 | | | |
| $\beta_{(1–6).4}$ | 2.056/normal | 0.007 | 0.003 | 0.135 | 0.134 | | | |
| $\beta_{(7–12).1}$ | 0.432/right-skewed | −0.000 | −0.001 | 0.066 | 0.066 | | | |
| $\beta_{(7–12).2}$ | 1.000/right-skewed | 0.001 | 0.001 | 0.083 | 0.083 | | | |
| $\beta_{(7–12).3}$ | 1.527/right-skewed | 0.004 | 0.003 | 0.109 | 0.109 | | | |
| $\beta_{(7–12).4}$ | 2.027/right-skewed | 0.007 | 0.003 | 0.144 | 0.144 | | | |

*Table E.12.* IRT-grm standardized parameters of interest and corresponding standard errors in Cell rnNS6 averaged over items with equal population value/shape. $n = 600$; $R = 1000$.

| | $\omega$ | PB($\hat{\omega}$) | RB($\hat{\omega}$) | RMSE($\hat{\omega}$) | SD($\hat{\omega}$) | RB($\hat{se}$) | RMSE($\hat{se}$) | 95%-cov. |
|---|---|---|---|---|---|---|---|---|
| $\lambda_{(1–6)}$ | 0.800/normal | −0.004 | −0.005 | 0.020 | 0.020 | −0.008 | 0.002 | 0.950 |
| $\lambda_{(7–12)}$ | 0.800/right-skewed | −0.008 | −0.010 | 0.024 | 0.023 | −0.008 | 0.002 | 0.945 |
| $\tau_{(1–6).1}$ | −1.645/normal | −0.002 | 0.001 | 0.090 | 0.090 | −0.017 | 0.007 | 0.946 |
| $\tau_{(1–6).2}$ | −0.643/normal | 0.012 | −0.019 | 0.055 | 0.054 | −0.000 | 0.002 | 0.945 |
| $\tau_{(1–6).3}$ | 0.643/normal | −0.010 | −0.016 | 0.055 | 0.054 | −0.010 | 0.001 | 0.945 |
| $\tau_{(1–6).4}$ | 1.645/normal | −0.004 | −0.002 | 0.089 | 0.089 | −0.018 | 0.006 | 0.946 |
| $\tau_{(7–12).1}$ | 0.346/right-skewed | −0.008 | −0.022 | 0.051 | 0.050 | 0.008 | 0.001 | 0.948 |
| $\tau_{(7–12).2}$ | 0.800/right-skewed | −0.017 | −0.021 | 0.058 | 0.056 | 0.001 | 0.002 | 0.939 |
| $\tau_{(7–12).3}$ | 1.221/right-skewed | −0.017 | −0.014 | 0.070 | 0.067 | −0.009 | 0.003 | 0.937 |
| $\tau_{(7–12).4}$ | 1.622/right-skewed | −0.007 | −0.004 | 0.088 | 0.088 | −0.020 | 0.006 | 0.941 |
| $\alpha_{(1–6)}$ | 2.269/normal | −0.021 | −0.009 | 0.156 | 0.155 | −0.011 | 0.011 | 0.944 |
| $\alpha_{(7–12)}$ | 2.269/right-skewed | −0.047 | −0.021 | 0.181 | 0.175 | −0.005 | 0.015 | 0.929 |
| $\beta_{(1–6).1}$ | −2.056/normal | −0.014 | 0.007 | 0.134 | 0.134 | | | |
| $\beta_{(1–6).2}$ | −0.804/normal | 0.011 | −0.013 | 0.073 | 0.072 | | | |
| $\beta_{(1–6).3}$ | 0.804/normal | −0.008 | −0.010 | 0.073 | 0.073 | | | |
| $\beta_{(1–6).4}$ | 2.056/normal | 0.008 | 0.004 | 0.135 | 0.134 | | | |
| $\beta_{(7–12).1}$ | 0.432/right-skewed | −0.005 | −0.012 | 0.064 | 0.064 | | | |
| $\beta_{(7–12).2}$ | 1.000/right-skewed | −0.010 | −0.010 | 0.080 | 0.079 | | | |
| $\beta_{(7–12).3}$ | 1.527/right-skewed | −0.005 | −0.003 | 0.105 | 0.105 | | | |
| $\beta_{(7–12).4}$ | 2.027/right-skewed | 0.014 | 0.007 | 0.142 | 0.141 | | | |

*Table E.13.* FA-lin standardized parameters of interest and corresponding standard errors in Cell lnNS2 averaged over items with equal population value/shape. $n = 200$; $R = 1000$.

| | $\omega$ | PB($\hat{\omega}$) | RB($\hat{\omega}$) | RMSE($\hat{\omega}$) | SD($\hat{\omega}$) | RB($\hat{se}$) | RMSE($\hat{se}$) | 95%-cov. |
|---|---|---|---|---|---|---|---|---|
| $\lambda_{(1–6)}$ | 0.800/normal | −0.063 | −0.078 | 0.072 | 0.035 | 0.017 | 0.004 | 0.611 |
| $\lambda_{(7–12)}$ | 0.800/left-skewed | −0.098 | −0.122 | 0.107 | 0.044 | −0.120 | 0.007 | 0.280 |

*Table E.14.* FA-poly standardized parameters of interest and corresponding standard errors in Cell lnNS2 averaged over items with equal population value/shape. $n = 200$; $R = 1000$.

| | $\omega$ | PB($\hat{\omega}$) | RB($\hat{\omega}$) | RMSE($\hat{\omega}$) | SD($\hat{\omega}$) | RB($\hat{se}$) | RMSE($\hat{se}$) | 95%-cov. |
|---|---|---|---|---|---|---|---|---|
| $\lambda_{(1–6)}$ | 0.800/normal | 0.002 | 0.002 | 0.034 | 0.034 | −0.056 | 0.005 | 0.925 |
| $\lambda_{(7–12)}$ | 0.800/left-skewed | 0.001 | 0.001 | 0.040 | 0.040 | −0.059 | 0.006 | 0.929 |
| $\tau_{(1–6).1}$ | −1.645/normal | −0.023 | 0.014 | 0.158 | 0.156 | −0.018 | 0.019 | 0.952 |
| $\tau_{(1–6).2}$ | −0.643/normal | −0.004 | 0.007 | 0.095 | 0.095 | 0.011 | 0.003 | 0.948 |
| $\tau_{(1–6).3}$ | 0.643/normal | 0.004 | 0.006 | 0.096 | 0.096 | −0.001 | 0.003 | 0.942 |
| $\tau_{(1–6).4}$ | 1.645/normal | 0.023 | 0.014 | 0.163 | 0.162 | −0.048 | 0.021 | 0.945 |
| $\tau_{(7–12).1}$ | −1.622/left-skewed | −0.021 | 0.013 | 0.155 | 0.154 | −0.019 | 0.017 | 0.951 |
| $\tau_{(7–12).2}$ | −1.221/left-skewed | −0.011 | 0.009 | 0.121 | 0.120 | −0.013 | 0.007 | 0.957 |
| $\tau_{(7–12).3}$ | −0.800/left-skewed | −0.003 | 0.004 | 0.100 | 0.100 | 0.002 | 0.004 | 0.945 |
| $\tau_{(7–12).4}$ | −0.346/left-skewed | 0.000 | −0.000 | 0.090 | 0.090 | 0.004 | 0.002 | 0.953 |
| $\alpha_{(1–6)}$ | 2.269/normal | 0.046 | 0.020 | 0.279 | 0.275 | −0.059 | 0.037 | 0.940 |
| $\alpha_{(7–12)}$ | 2.269/left-skewed | 0.053 | 0.024 | 0.337 | 0.333 | −0.062 | 0.058 | 0.945 |
| $\beta_{(1–6).1}$ | −2.056/normal | −0.030 | 0.014 | 0.239 | 0.237 | | | |
| $\beta_{(1–6).2}$ | −0.804/normal | −0.005 | 0.007 | 0.127 | 0.127 | | | |
| $\beta_{(1–6).3}$ | 0.804/normal | 0.005 | 0.006 | 0.130 | 0.130 | | | |
| $\beta_{(1–6).4}$ | 2.056/normal | 0.030 | 0.014 | 0.241 | 0.239 | | | |
| $\beta_{(7–12).1}$ | −2.027/left-skewed | −0.032 | 0.016 | 0.253 | 0.251 | | | |
| $\beta_{(7–12).2}$ | −1.527/left-skewed | −0.017 | 0.011 | 0.191 | 0.190 | | | |
| $\beta_{(7–12).3}$ | −1.000/left-skewed | −0.006 | 0.006 | 0.143 | 0.143 | | | |
| $\beta_{(7–12).4}$ | −0.432/left-skewed | −0.000 | 0.001 | 0.115 | 0.115 | | | |

*Table E.15.* IRT-grm standardized parameters of interest and corresponding standard errors in Cell lnNS2 averaged over items with equal population value/shape. $n = 200$; $R = 1000$.

| | $\omega$ | PB($\hat{\omega}$) | RB($\hat{\omega}$) | RMSE($\hat{\omega}$) | SD($\hat{\omega}$) | RB($\hat{se}$) | RMSE($\hat{se}$) | 95%-cov. |
|---|---|---|---|---|---|---|---|---|
| $\lambda_{(1–6)}$ | 0.800/normal | −0.004 | −0.005 | 0.035 | 0.035 | −0.010 | 0.005 | 0.943 |
| $\lambda_{(7–12)}$ | 0.800/left-skewed | −0.009 | −0.011 | 0.041 | 0.040 | 0.006 | 0.006 | 0.952 |
| $\tau_{(1–6).1}$ | −1.645/normal | −0.012 | 0.007 | 0.158 | 0.157 | −0.013 | 0.021 | 0.956 |
| $\tau_{(1–6).2}$ | −0.643/normal | 0.008 | −0.013 | 0.092 | 0.092 | 0.012 | 0.004 | 0.951 |
| $\tau_{(1–6).3}$ | 0.643/normal | −0.008 | −0.012 | 0.095 | 0.095 | −0.010 | 0.004 | 0.942 |
| $\tau_{(1–6).4}$ | 1.645/normal | 0.021 | 0.013 | 0.165 | 0.164 | −0.047 | 0.023 | 0.951 |
| $\tau_{(7–12).1}$ | −1.622/left-skewed | −0.007 | 0.005 | 0.155 | 0.154 | −0.014 | 0.020 | 0.957 |
| $\tau_{(7–12).2}$ | −1.221/left-skewed | 0.012 | −0.010 | 0.118 | 0.118 | −0.011 | 0.009 | 0.945 |
| $\tau_{(7–12).3}$ | −0.800/left-skewed | 0.017 | −0.021 | 0.097 | 0.096 | 0.010 | 0.005 | 0.948 |
| $\tau_{(7–12).4}$ | −0.346/left-skewed | 0.008 | −0.024 | 0.088 | 0.087 | 0.006 | 0.003 | 0.951 |
| $\alpha_{(1–6)}$ | 2.269/normal | −0.001 | −0.000 | 0.274 | 0.274 | −0.014 | 0.034 | 0.943 |
| $\alpha_{(7–12)}$ | 2.269/left-skewed | −0.029 | −0.013 | 0.310 | 0.308 | −0.000 | 0.047 | 0.937 |
| $\beta_{(1–6).1}$ | −2.056/normal | −0.032 | 0.016 | 0.241 | 0.239 | | | |
| $\beta_{(1–6).2}$ | −0.804/normal | 0.004 | −0.005 | 0.124 | 0.123 | | | |
| $\beta_{(1–6).3}$ | 0.804/normal | −0.004 | −0.005 | 0.127 | 0.127 | | | |
| $\beta_{(1–6).4}$ | 2.056/normal | 0.043 | 0.021 | 0.249 | 0.246 | | | |
| $\beta_{(7–12).1}$ | −2.027/left-skewed | −0.040 | 0.020 | 0.252 | 0.248 | | | |
| $\beta_{(7–12).2}$ | −1.527/left-skewed | −0.007 | 0.005 | 0.183 | 0.183 | | | |
| $\beta_{(7–12).3}$ | −1.000/left-skewed | 0.007 | −0.007 | 0.136 | 0.136 | | | |
| $\beta_{(7–12).4}$ | −0.432/left-skewed | 0.005 | −0.012 | 0.111 | 0.111 | | | |

*Table E.16.* FA-lin standardized parameters of interest and corresponding standard errors in Cell lnNS6 averaged over items with equal population value/shape. $n = 600$; $R = 1000$.

| | $\omega$ | PB($\hat{\omega}$) | RB($\hat{\omega}$) | RMSE($\hat{\omega}$) | SD($\hat{\omega}$) | RB($\hat{se}$) | RMSE($\hat{se}$) | 95%-cov. |
|---|---|---|---|---|---|---|---|---|
| $\lambda_{(1–6)}$ | 0.800/normal | −0.061 | −0.076 | 0.064 | 0.020 | 0.022 | 0.001 | 0.109 |
| $\lambda_{(7–12)}$ | 0.800/left-skewed | −0.095 | −0.119 | 0.099 | 0.026 | −0.137 | 0.004 | 0.009 |

*Table E.17.* FA-poly standardized parameters of interest and corresponding standard errors in Cell lnNS6 averaged over items with equal population value/shape. $n = 600$; $R = 1000$.

| | $\omega$ | PB($\hat{\omega}$) | RB($\hat{\omega}$) | RMSE($\hat{\omega}$) | SD($\hat{\omega}$) | RB($\hat{se}$) | RMSE($\hat{se}$) | 95%-cov. |
|---|---|---|---|---|---|---|---|---|
| $\lambda_{(1–6)}$ | 0.800/normal | 0.001 | 0.001 | 0.019 | 0.019 | −0.026 | 0.002 | 0.940 |
| $\lambda_{(7–12)}$ | 0.800/left-skewed | 0.002 | 0.002 | 0.024 | 0.024 | −0.046 | 0.002 | 0.927 |
| $\tau_{(1–6).1}$ | −1.645/normal | −0.005 | 0.003 | 0.088 | 0.088 | −0.008 | 0.005 | 0.952 |
| $\tau_{(1–6).2}$ | −0.643/normal | −0.001 | 0.002 | 0.055 | 0.055 | 0.007 | 0.001 | 0.952 |
| $\tau_{(1–6).3}$ | 0.643/normal | −0.000 | −0.000 | 0.055 | 0.055 | 0.012 | 0.001 | 0.956 |
| $\tau_{(1–6).4}$ | 1.645/normal | 0.007 | 0.004 | 0.089 | 0.089 | −0.024 | 0.006 | 0.949 |
| $\tau_{(7–12).1}$ | −1.622/left-skewed | −0.005 | 0.003 | 0.087 | 0.087 | −0.018 | 0.005 | 0.946 |
| $\tau_{(7–12).2}$ | −1.221/left-skewed | −0.003 | 0.003 | 0.068 | 0.068 | 0.002 | 0.003 | 0.950 |
| $\tau_{(7–12).3}$ | −0.800/left-skewed | −0.002 | 0.002 | 0.057 | 0.057 | 0.009 | 0.001 | 0.952 |
| $\tau_{(7–12).4}$ | −0.346/left-skewed | −0.002 | 0.006 | 0.052 | 0.052 | 0.014 | 0.001 | 0.956 |
| $\alpha_{(1–6)}$ | 2.269/normal | 0.019 | 0.008 | 0.156 | 0.155 | −0.026 | 0.011 | 0.947 |
| $\alpha_{(7–12)}$ | 2.269/left-skewed | 0.028 | 0.012 | 0.193 | 0.191 | −0.045 | 0.019 | 0.940 |
| $\beta_{(1–6).1}$ | −2.056/normal | −0.005 | 0.002 | 0.132 | 0.132 | | | |
| $\beta_{(1–6).2}$ | −0.804/normal | −0.001 | 0.001 | 0.074 | 0.074 | | | |
| $\beta_{(1–6).3}$ | 0.804/normal | −0.001 | −0.001 | 0.074 | 0.074 | | | |
| $\beta_{(1–6).4}$ | 2.056/normal | 0.008 | 0.004 | 0.133 | 0.132 | | | |
| $\beta_{(7–12).1}$ | −2.027/left-skewed | −0.006 | 0.003 | 0.143 | 0.142 | | | |
| $\beta_{(7–12).2}$ | −1.527/left-skewed | −0.003 | 0.002 | 0.108 | 0.108 | | | |
| $\beta_{(7–12).3}$ | −1.000/left-skewed | −0.001 | 0.001 | 0.081 | 0.081 | | | |
| $\beta_{(7–12).4}$ | −0.432/left-skewed | −0.002 | 0.005 | 0.065 | 0.065 | | | |

*Table E.18.* IRT-grm standardized parameters of interest and corresponding standard errors in Cell lnNS6 averaged over items with equal population value/shape. $n = 600$; $R = 1000$.

| | $\omega$ | PB($\hat{\omega}$) | RB($\hat{\omega}$) | RMSE($\hat{\omega}$) | SD($\hat{\omega}$) | RB($\hat{se}$) | RMSE($\hat{se}$) | 95%-cov. |
|---|---|---|---|---|---|---|---|---|
| $\lambda_{(1–6)}$ | 0.800/normal | −0.004 | −0.005 | 0.020 | 0.020 | −0.009 | 0.001 | 0.951 |
| $\lambda_{(7–12)}$ | 0.800/left-skewed | −0.007 | −0.009 | 0.025 | 0.023 | −0.020 | 0.002 | 0.944 |
| $\tau_{(1–6).1}$ | −1.645/normal | 0.004 | −0.003 | 0.088 | 0.088 | −0.004 | 0.006 | 0.947 |
| $\tau_{(1–6).2}$ | −0.643/normal | 0.010 | −0.016 | 0.054 | 0.053 | 0.011 | 0.001 | 0.947 |
| $\tau_{(1–6).3}$ | 0.643/normal | −0.012 | −0.018 | 0.054 | 0.053 | 0.014 | 0.001 | 0.947 |
| $\tau_{(1–6).4}$ | 1.645/normal | 0.003 | 0.002 | 0.090 | 0.090 | −0.020 | 0.006 | 0.948 |
| $\tau_{(7–12).1}$ | −1.622/left-skewed | 0.006 | −0.004 | 0.088 | 0.087 | −0.015 | 0.006 | 0.944 |
| $\tau_{(7–12).2}$ | −1.221/left-skewed | 0.018 | −0.014 | 0.068 | 0.066 | 0.009 | 0.003 | 0.942 |
| $\tau_{(7–12).3}$ | −0.800/left-skewed | 0.016 | −0.020 | 0.057 | 0.055 | 0.014 | 0.002 | 0.938 |
| $\tau_{(7–12).4}$ | −0.346/left-skewed | 0.005 | −0.016 | 0.050 | 0.050 | 0.021 | 0.001 | 0.953 |
| $\alpha_{(1–6)}$ | 2.269/normal | −0.020 | −0.009 | 0.156 | 0.154 | −0.010 | 0.011 | 0.945 |
| $\alpha_{(7–12)}$ | 2.269/left-skewed | −0.044 | −0.019 | 0.183 | 0.177 | −0.017 | 0.015 | 0.926 |
| $\beta_{(1–6).1}$ | −2.056/normal | −0.006 | 0.003 | 0.133 | 0.132 | | | |
| $\beta_{(1–6).2}$ | −0.804/normal | 0.008 | −0.010 | 0.072 | 0.071 | | | |
| $\beta_{(1–6).3}$ | 0.804/normal | −0.010 | −0.013 | 0.072 | 0.071 | | | |
| $\beta_{(1–6).4}$ | 2.056/normal | 0.015 | 0.007 | 0.135 | 0.135 | | | |
| $\beta_{(7–12).1}$ | −2.027/left-skewed | −0.014 | 0.007 | 0.141 | 0.141 | | | |
| $\beta_{(7–12).2}$ | −1.527/left-skewed | 0.006 | −0.004 | 0.104 | 0.103 | | | |
| $\beta_{(7–12).3}$ | −1.000/left-skewed | 0.010 | −0.010 | 0.078 | 0.078 | | | |
| $\beta_{(7–12).4}$ | −0.432/left-skewed | 0.002 | −0.006 | 0.063 | 0.063 | | | |

*Table E.19.* FA-lin standardized parameters of interest and corresponding standard errors in Cell lrnNS2 averaged over items with equal population value/shape. $n = 200$; $R = 1000$.

| | $\omega$ | PB($\hat{\omega}$) | RB($\hat{\omega}$) | RMSE($\hat{\omega}$) | SD($\hat{\omega}$) | RB($\hat{se}$) | RMSE($\hat{se}$) | 95%-cov. |
|---|---|---|---|---|---|---|---|---|
| $\lambda_{(1–4)}$ | 0.800/normal | −0.042 | −0.052 | 0.054 | 0.034 | −0.010 | 0.004 | 0.825 |
| $\lambda_{(5–8)}$ | 0.800/left-skewed | −0.153 | −0.192 | 0.160 | 0.046 | −0.029 | 0.004 | 0.033 |
| $\lambda_{(9–12)}$ | 0.800/right-skewed | −0.152 | −0.190 | 0.159 | 0.046 | −0.041 | 0.005 | 0.038 |

*Table E.20.* FA-poly standardized parameters of interest and corresponding standard errors in Cell lrnNS2 averaged over items with equal population value/shape. $n = 200$; $R = 1000$.

| | $\omega$ | PB($\hat{\omega}$) | RB($\hat{\omega}$) | RMSE($\hat{\omega}$) | SD($\hat{\omega}$) | RB($\hat{se}$) | RMSE($\hat{se}$) | 95%-cov. |
|---|---|---|---|---|---|---|---|---|
| $\lambda_{(1–4)}$ | 0.800/normal | 0.002 | 0.002 | 0.033 | 0.033 | −0.056 | 0.004 | 0.930 |
| $\lambda_{(5–8)}$ | 0.800/left-skewed | 0.003 | 0.004 | 0.042 | 0.042 | −0.075 | 0.006 | 0.918 |
| $\lambda_{(9–12)}$ | 0.800/right-skewed | 0.003 | 0.004 | 0.043 | 0.043 | −0.091 | 0.006 | 0.917 |
| $\tau_{(1–4).1}$ | −1.645/normal | −0.022 | 0.013 | 0.160 | 0.159 | −0.032 | 0.019 | 0.950 |
| $\tau_{(1–4).2}$ | −0.643/normal | −0.004 | 0.006 | 0.098 | 0.098 | −0.022 | 0.003 | 0.940 |
| $\tau_{(1–4).3}$ | 0.643/normal | 0.004 | 0.007 | 0.097 | 0.097 | −0.006 | 0.003 | 0.944 |
| $\tau_{(1–4).4}$ | 1.645/normal | 0.017 | 0.010 | 0.162 | 0.161 | −0.046 | 0.020 | 0.949 |
| $\tau_{(5–8).1}$ | −1.622/left-skewed | −0.021 | 0.013 | 0.160 | 0.158 | −0.045 | 0.019 | 0.944 |
| $\tau_{(5–8).2}$ | −1.221/left-skewed | −0.011 | 0.009 | 0.123 | 0.122 | −0.031 | 0.008 | 0.954 |
| $\tau_{(5–8).3}$ | −0.800/left-skewed | −0.006 | 0.008 | 0.103 | 0.103 | −0.026 | 0.004 | 0.942 |
| $\tau_{(5–8).4}$ | −0.346/left-skewed | −0.002 | 0.005 | 0.092 | 0.092 | −0.012 | 0.003 | 0.950 |
| $\tau_{(9–12).1}$ | 0.346/right-skewed | 0.001 | 0.004 | 0.091 | 0.091 | 0.001 | 0.002 | 0.956 |
| $\tau_{(9–12).2}$ | 0.800/right-skewed | 0.005 | 0.006 | 0.099 | 0.099 | 0.008 | 0.004 | 0.947 |
| $\tau_{(9–12).3}$ | 1.221/right-skewed | 0.010 | 0.008 | 0.122 | 0.121 | −0.021 | 0.008 | 0.958 |
| $\tau_{(9–12).4}$ | 1.622/right-skewed | 0.019 | 0.011 | 0.157 | 0.156 | −0.033 | 0.018 | 0.946 |
| $\alpha_{(1–4)}$ | 2.269/normal | 0.045 | 0.020 | 0.273 | 0.269 | −0.049 | 0.035 | 0.939 |
| $\alpha_{(5–8)}$ | 2.269/left-skewed | 0.074 | 0.033 | 0.366 | 0.358 | −0.089 | 0.076 | 0.944 |
| $\alpha_{(9–12)}$ | 2.269/right-skewed | 0.078 | 0.034 | 0.374 | 0.366 | −0.106 | 0.081 | 0.940 |
| $\beta_{(1–4).1}$ | −2.056/normal | −0.028 | 0.014 | 0.241 | 0.239 | | | |
| $\beta_{(1–4).2}$ | −0.804/normal | −0.005 | 0.006 | 0.132 | 0.132 | | | |
| $\beta_{(1–4).3}$ | 0.804/normal | 0.006 | 0.007 | 0.129 | 0.129 | | | |
| $\beta_{(1–4).4}$ | 2.056/normal | 0.022 | 0.011 | 0.240 | 0.239 | | | |
| $\beta_{(5–8).1}$ | −2.027/left-skewed | −0.027 | 0.013 | 0.257 | 0.255 | | | |
| $\beta_{(5–8).2}$ | −1.527/left-skewed | −0.014 | 0.009 | 0.194 | 0.193 | | | |
| $\beta_{(5–8).3}$ | −1.000/left-skewed | −0.008 | 0.008 | 0.150 | 0.149 | | | |
| $\beta_{(5–8).4}$ | −0.432/left-skewed | −0.002 | 0.004 | 0.117 | 0.118 | | | |
| $\beta_{(9–12).1}$ | 0.432/right-skewed | 0.001 | 0.003 | 0.115 | 0.115 | | | |
| $\beta_{(9–12).2}$ | 1.000/right-skewed | 0.006 | 0.006 | 0.144 | 0.144 | | | |
| $\beta_{(9–12).3}$ | 1.527/right-skewed | 0.013 | 0.009 | 0.195 | 0.195 | | | |
| $\beta_{(9–12).4}$ | 2.027/right-skewed | 0.024 | 0.012 | 0.257 | 0.256 | | | |

*Table E.21.* IRT-grm standardized parameters of interest and corresponding standard errors in Cell lrnNS2 averaged over items with equal population value/shape. $n = 200$; $R = 1000$.

| | $\omega$ | PB($\hat{\omega}$) | RB($\hat{\omega}$) | RMSE($\hat{\omega}$) | SD($\hat{\omega}$) | RB($\hat{se}$) | RMSE($\hat{se}$) | 95%-cov. |
|---|---|---|---|---|---|---|---|---|
| $\lambda_{(1–4)}$ | 0.800/normal | −0.004 | −0.005 | 0.035 | 0.035 | −0.008 | 0.004 | 0.950 |
| $\lambda_{(5–8)}$ | 0.800/left-skewed | −0.009 | −0.011 | 0.043 | 0.042 | −0.027 | 0.006 | 0.942 |
| $\lambda_{(9–12)}$ | 0.800/right-skewed | −0.009 | −0.011 | 0.043 | 0.042 | −0.021 | 0.006 | 0.948 |
| $\tau_{(1–4).1}$ | −1.645/normal | −0.015 | 0.009 | 0.163 | 0.162 | −0.042 | 0.022 | 0.948 |
| $\tau_{(1–4).2}$ | −0.643/normal | 0.009 | −0.014 | 0.096 | 0.096 | −0.026 | 0.004 | 0.944 |
| $\tau_{(1–4).3}$ | 0.643/normal | −0.009 | −0.013 | 0.095 | 0.095 | −0.014 | 0.004 | 0.945 |
| $\tau_{(1–4).4}$ | 1.645/normal | 0.012 | 0.007 | 0.164 | 0.164 | −0.052 | 0.023 | 0.952 |
| $\tau_{(5–8).1}$ | −1.622/left-skewed | −0.012 | 0.007 | 0.161 | 0.161 | −0.051 | 0.023 | 0.950 |
| $\tau_{(5–8).2}$ | −1.221/left-skewed | 0.010 | −0.008 | 0.122 | 0.122 | −0.040 | 0.010 | 0.943 |
| $\tau_{(5–8).3}$ | −0.800/left-skewed | 0.013 | −0.017 | 0.101 | 0.100 | −0.029 | 0.005 | 0.941 |
| $\tau_{(5–8).4}$ | −0.346/left-skewed | 0.008 | −0.023 | 0.090 | 0.090 | −0.017 | 0.003 | 0.948 |
| $\tau_{(9–12).1}$ | 0.346/right-skewed | −0.009 | −0.025 | 0.089 | 0.089 | −0.012 | 0.003 | 0.946 |
| $\tau_{(9–12).2}$ | 0.800/right-skewed | −0.015 | −0.019 | 0.099 | 0.098 | −0.005 | 0.005 | 0.943 |
| $\tau_{(9–12).3}$ | 1.221/right-skewed | −0.010 | −0.008 | 0.120 | 0.120 | −0.024 | 0.010 | 0.944 |
| $\tau_{(9–12).4}$ | 1.622/right-skewed | 0.010 | 0.006 | 0.158 | 0.158 | −0.034 | 0.021 | 0.950 |
| $\alpha_{(1–4)}$ | 2.269/normal | −0.002 | −0.001 | 0.271 | 0.271 | −0.002 | 0.034 | 0.941 |
| $\alpha_{(5–8)}$ | 2.269/left-skewed | −0.024 | −0.011 | 0.328 | 0.327 | −0.040 | 0.053 | 0.932 |
| $\alpha_{(9–12)}$ | 2.269/right-skewed | −0.026 | −0.011 | 0.323 | 0.322 | −0.028 | 0.049 | 0.936 |
| $\beta_{(1–4).1}$ | −2.056/normal | −0.036 | 0.018 | 0.250 | 0.247 | | | |
| $\beta_{(1–4).2}$ | −0.804/normal | 0.005 | −0.006 | 0.129 | 0.129 | | | |
| $\beta_{(1–4).3}$ | 0.804/normal | −0.005 | −0.006 | 0.127 | 0.127 | | | |
| $\beta_{(1–4).4}$ | 2.056/normal | 0.032 | 0.015 | 0.248 | 0.246 | | | |
| $\beta_{(5–8).1}$ | −2.027/left-skewed | −0.046 | 0.023 | 0.262 | 0.258 | | | |
| $\beta_{(5–8).2}$ | −1.527/left-skewed | −0.010 | 0.007 | 0.190 | 0.190 | | | |
| $\beta_{(5–8).3}$ | −1.000/left-skewed | 0.002 | −0.002 | 0.144 | 0.144 | | | |
| $\beta_{(5–8).4}$ | −0.432/left-skewed | 0.004 | −0.010 | 0.115 | 0.115 | | | |
| $\beta_{(9–12).1}$ | 0.432/right-skewed | −0.005 | −0.012 | 0.114 | 0.114 | | | |
| $\beta_{(9–12).2}$ | 1.000/right-skewed | −0.004 | −0.004 | 0.139 | 0.139 | | | |
| $\beta_{(9–12).3}$ | 1.527/right-skewed | 0.010 | 0.007 | 0.190 | 0.189 | | | |
| $\beta_{(9–12).4}$ | 2.027/right-skewed | 0.044 | 0.022 | 0.260 | 0.256 | | | |

*Table E.22.* FA-lin standardized parameters of interest and corresponding standard errors in Cell lrnNS6 averaged over items with equal population value/shape. $n = 600$; $R = 1000$.

| | $\omega$ | PB($\hat{\omega}$) | RB($\hat{\omega}$) | RMSE($\hat{\omega}$) | SD($\hat{\omega}$) | RB($\hat{se}$) | RMSE($\hat{se}$) | 95%-cov. |
|---|---|---|---|---|---|---|---|---|
| $\lambda_{(1–4)}$ | 0.800/normal | −0.041 | −0.051 | 0.045 | 0.019 | −0.002 | 0.001 | 0.452 |
| $\lambda_{(5–8)}$ | 0.800/left-skewed | −0.150 | −0.188 | 0.152 | 0.026 | −0.022 | 0.002 | 0.000 |
| $\lambda_{(9–12)}$ | 0.800/right-skewed | −0.152 | −0.189 | 0.154 | 0.026 | −0.004 | 0.001 | 0.000 |

*Table E.23.* FA-poly standardized parameters of interest and corresponding standard errors in Cell lrnNS6 averaged over items with equal population value/shape. $n = 600$; $R = 1000$.

| | $\omega$ | PB($\hat{\omega}$) | RB($\hat{\omega}$) | RMSE($\hat{\omega}$) | SD($\hat{\omega}$) | RB($\hat{se}$) | RMSE($\hat{se}$) | 95%-cov. |
|---|---|---|---|---|---|---|---|---|
| $\lambda_{(1–4)}$ | 0.800/normal | 0.001 | 0.001 | 0.019 | 0.019 | −0.019 | 0.001 | 0.943 |
| $\lambda_{(5–8)}$ | 0.800/left-skewed | 0.003 | 0.003 | 0.024 | 0.024 | −0.021 | 0.002 | 0.932 |
| $\lambda_{(9–12)}$ | 0.800/right-skewed | 0.001 | 0.002 | 0.023 | 0.023 | −0.009 | 0.002 | 0.942 |
| $\tau_{(1–4).1}$ | −1.645/normal | −0.007 | 0.004 | 0.089 | 0.089 | −0.017 | 0.006 | 0.946 |
| $\tau_{(1–4).2}$ | −0.643/normal | −0.003 | 0.004 | 0.056 | 0.056 | −0.011 | 0.001 | 0.946 |
| $\tau_{(1–4).3}$ | 0.643/normal | −0.000 | −0.000 | 0.055 | 0.055 | −0.001 | 0.001 | 0.947 |
| $\tau_{(1–4).4}$ | 1.645/normal | 0.007 | 0.004 | 0.087 | 0.087 | 0.003 | 0.005 | 0.954 |
| $\tau_{(5–8).1}$ | −1.622/left-skewed | −0.004 | 0.003 | 0.087 | 0.087 | −0.017 | 0.005 | 0.944 |
| $\tau_{(5–8).2}$ | −1.221/left-skewed | −0.003 | 0.003 | 0.068 | 0.068 | 0.000 | 0.002 | 0.948 |
| $\tau_{(5–8).3}$ | −0.800/left-skewed | −0.003 | 0.003 | 0.057 | 0.057 | 0.004 | 0.001 | 0.953 |
| $\tau_{(5–8).4}$ | −0.346/left-skewed | −0.002 | 0.006 | 0.052 | 0.052 | 0.007 | 0.001 | 0.954 |
| $\tau_{(9–12).1}$ | 0.346/right-skewed | −0.002 | −0.005 | 0.052 | 0.052 | −0.002 | 0.001 | 0.955 |
| $\tau_{(9–12).2}$ | 0.800/right-skewed | −0.001 | −0.001 | 0.056 | 0.057 | 0.020 | 0.002 | 0.952 |
| $\tau_{(9–12).3}$ | 1.221/right-skewed | 0.001 | 0.001 | 0.067 | 0.067 | 0.016 | 0.003 | 0.951 |
| $\tau_{(9–12).4}$ | 1.622/right-skewed | 0.003 | 0.002 | 0.084 | 0.084 | 0.017 | 0.005 | 0.953 |
| $\alpha_{(1–4)}$ | 2.269/normal | 0.014 | 0.006 | 0.154 | 0.153 | −0.022 | 0.011 | 0.947 |
| $\alpha_{(5–8)}$ | 2.269/left-skewed | 0.035 | 0.016 | 0.197 | 0.193 | −0.028 | 0.021 | 0.950 |
| $\alpha_{(9–12)}$ | 2.269/right-skewed | 0.024 | 0.011 | 0.190 | 0.188 | −0.009 | 0.020 | 0.946 |
| $\beta_{(1–4).1}$ | −2.056/normal | −0.009 | 0.004 | 0.132 | 0.132 | | | |
| $\beta_{(1–4).2}$ | −0.804/normal | −0.004 | 0.005 | 0.075 | 0.075 | | | |
| $\beta_{(1–4).3}$ | 0.804/normal | −0.000 | −0.000 | 0.074 | 0.074 | | | |
| $\beta_{(1–4).4}$ | 2.056/normal | 0.008 | 0.004 | 0.130 | 0.130 | | | |
| $\beta_{(5–8).1}$ | −2.027/left-skewed | −0.002 | 0.001 | 0.140 | 0.140 | | | |
| $\beta_{(5–8).2}$ | −1.527/left-skewed | −0.001 | 0.001 | 0.107 | 0.107 | | | |
| $\beta_{(5–8).3}$ | −1.000/left-skewed | −0.002 | 0.002 | 0.082 | 0.082 | | | |
| $\beta_{(5–8).4}$ | −0.432/left-skewed | −0.002 | 0.004 | 0.066 | 0.066 | | | |
| $\beta_{(9–12).1}$ | 0.432/right-skewed | −0.003 | −0.006 | 0.067 | 0.067 | | | |
| $\beta_{(9–12).2}$ | 1.000/right-skewed | −0.001 | −0.001 | 0.081 | 0.081 | | | |
| $\beta_{(9–12).3}$ | 1.527/right-skewed | 0.001 | 0.000 | 0.106 | 0.106 | | | |
| $\beta_{(9–12).4}$ | 2.027/right-skewed | 0.004 | 0.002 | 0.136 | 0.136 | | | |

*Table E.24.* IRT-grm standardized parameters of interest and corresponding standard errors in Cell lrnNS6 averaged over items with equal population value/shape. $n = 600$; $R = 1000$.

| | $\omega$ | PB($\hat{\omega}$) | RB($\hat{\omega}$) | RMSE($\hat{\omega}$) | SD($\hat{\omega}$) | RB($\hat{se}$) | RMSE($\hat{se}$) | 95%-cov. |
|---|---|---|---|---|---|---|---|---|
| $\lambda_{(1–4)}$ | 0.800/normal | −0.005 | −0.006 | 0.020 | 0.020 | −0.004 | 0.001 | 0.947 |
| $\lambda_{(5–8)}$ | 0.800/left-skewed | −0.007 | −0.009 | 0.024 | 0.023 | 0.000 | 0.002 | 0.947 |
| $\lambda_{(9–12)}$ | 0.800/right-skewed | −0.008 | −0.010 | 0.025 | 0.023 | 0.007 | 0.002 | 0.948 |
| $\tau_{(1–4).1}$ | −1.645/normal | −0.001 | 0.001 | 0.089 | 0.089 | −0.010 | 0.006 | 0.950 |
| $\tau_{(1–4).2}$ | −0.643/normal | 0.010 | −0.015 | 0.055 | 0.054 | −0.009 | 0.001 | 0.945 |
| $\tau_{(1–4).3}$ | 0.643/normal | −0.013 | −0.020 | 0.055 | 0.053 | 0.008 | 0.001 | 0.941 |
| $\tau_{(1–4).4}$ | 1.645/normal | −0.000 | −0.000 | 0.088 | 0.088 | −0.005 | 0.006 | 0.953 |
| $\tau_{(5–8).1}$ | −1.622/left-skewed | 0.005 | −0.003 | 0.087 | 0.087 | −0.013 | 0.006 | 0.948 |
| $\tau_{(5–8).2}$ | −1.221/left-skewed | 0.016 | −0.013 | 0.069 | 0.067 | −0.000 | 0.003 | 0.939 |
| $\tau_{(5–8).3}$ | −0.800/left-skewed | 0.016 | −0.020 | 0.058 | 0.056 | 0.004 | 0.001 | 0.941 |
| $\tau_{(5–8).4}$ | −0.346/left-skewed | 0.008 | −0.022 | 0.051 | 0.050 | 0.014 | 0.001 | 0.948 |
| $\tau_{(9–12).1}$ | 0.346/right-skewed | −0.011 | −0.032 | 0.052 | 0.051 | 0.002 | 0.001 | 0.944 |
| $\tau_{(9–12).2}$ | 0.800/right-skewed | −0.019 | −0.024 | 0.058 | 0.054 | 0.025 | 0.002 | 0.936 |
| $\tau_{(9–12).3}$ | 1.221/right-skewed | −0.019 | −0.016 | 0.068 | 0.065 | 0.021 | 0.003 | 0.939 |
| $\tau_{(9–12).4}$ | 1.622/right-skewed | −0.006 | −0.004 | 0.085 | 0.085 | 0.018 | 0.006 | 0.949 |
| $\alpha_{(1–4)}$ | 2.269/normal | −0.026 | −0.011 | 0.156 | 0.154 | −0.007 | 0.011 | 0.934 |
| $\alpha_{(5–8)}$ | 2.269/left-skewed | −0.043 | −0.019 | 0.183 | 0.178 | −0.002 | 0.016 | 0.931 |
| $\alpha_{(9–12)}$ | 2.269/right-skewed | −0.049 | −0.021 | 0.182 | 0.175 | 0.009 | 0.015 | 0.931 |
| $\beta_{(1–4).1}$ | −2.056/normal | −0.015 | 0.007 | 0.134 | 0.133 | | | |
| $\beta_{(1–4).2}$ | −0.804/normal | 0.007 | −0.009 | 0.073 | 0.073 | | | |
| $\beta_{(1–4).3}$ | 0.804/normal | −0.011 | −0.013 | 0.072 | 0.071 | | | |
| $\beta_{(1–4).4}$ | 2.056/normal | 0.013 | 0.006 | 0.133 | 0.133 | | | |
| $\beta_{(5–8).1}$ | −2.027/left-skewed | −0.016 | 0.008 | 0.140 | 0.139 | | | |
| $\beta_{(5–8).2}$ | −1.527/left-skewed | 0.004 | −0.003 | 0.104 | 0.104 | | | |
| $\beta_{(5–8).3}$ | −1.000/left-skewed | 0.009 | −0.009 | 0.080 | 0.079 | | | |
| $\beta_{(5–8).4}$ | −0.432/left-skewed | 0.005 | −0.012 | 0.064 | 0.064 | | | |
| $\beta_{(9–12).1}$ | 0.432/right-skewed | −0.009 | −0.021 | 0.065 | 0.064 | | | |
| $\beta_{(9–12).2}$ | 1.000/right-skewed | −0.013 | −0.013 | 0.078 | 0.077 | | | |
| $\beta_{(9–12).3}$ | 1.527/right-skewed | −0.007 | −0.005 | 0.102 | 0.102 | | | |
| $\beta_{(9–12).4}$ | 2.027/right-skewed | 0.015 | 0.007 | 0.137 | 0.136 | | | |

*Table E.25.* FA-lin standardized parameters of interest and corresponding standard errors in Cell bnNS2 averaged over items with equal population value/shape. $n = 200$; $R = 1000$.

| | $\omega$ | PB($\hat{\omega}$) | RB($\hat{\omega}$) | RMSE($\hat{\omega}$) | SD($\hat{\omega}$) | RB($\hat{se}$) | RMSE($\hat{se}$) | 95%-cov. |
|---|---|---|---|---|---|---|---|---|
| $\lambda_{(1–6)}$ | 0.800/normal | −0.047 | −0.059 | 0.057 | 0.033 | 0.005 | 0.004 | 0.763 |
| $\lambda_{(7–12)}$ | 0.800/bimodal | −0.049 | −0.062 | 0.060 | 0.033 | −0.003 | 0.004 | 0.738 |

*Table E.26.* FA-poly standardized parameters of interest and corresponding standard errors in Cell bnNS2 averaged over items with equal population value/shape. $n = 200$; $R = 1000$.

| | $\omega$ | PB($\hat{\omega}$) | RB($\hat{\omega}$) | RMSE($\hat{\omega}$) | SD($\hat{\omega}$) | RB($\hat{se}$) | RMSE($\hat{se}$) | 95%-cov. |
|---|---|---|---|---|---|---|---|---|
| $\lambda_{(1–6)}$ | 0.800/normal | 0.002 | 0.003 | 0.033 | 0.033 | −0.052 | 0.004 | 0.922 |
| $\lambda_{(7–12)}$ | 0.800/bimodal | 0.003 | 0.004 | 0.033 | 0.033 | −0.056 | 0.004 | 0.922 |
| $\tau_{(1–6).1}$ | −1.645/normal | −0.018 | 0.011 | 0.159 | 0.158 | −0.030 | 0.019 | 0.949 |
| $\tau_{(1–6).2}$ | −0.643/normal | −0.003 | 0.005 | 0.095 | 0.095 | 0.008 | 0.003 | 0.952 |
| $\tau_{(1–6).3}$ | 0.643/normal | 0.005 | 0.008 | 0.097 | 0.096 | −0.005 | 0.003 | 0.946 |
| $\tau_{(1–6).4}$ | 1.645/normal | 0.020 | 0.012 | 0.159 | 0.158 | −0.028 | 0.019 | 0.951 |
| $\tau_{(7–12).1}$ | −1.282/bimodal | −0.010 | 0.008 | 0.124 | 0.124 | −0.012 | 0.008 | 0.957 |
| $\tau_{(7–12).2}$ | −0.126/bimodal | 0.001 | −0.008 | 0.089 | 0.089 | −0.005 | 0.001 | 0.944 |
| $\tau_{(7–12).3}$ | 0.126/bimodal | 0.002 | 0.015 | 0.090 | 0.090 | −0.006 | 0.002 | 0.945 |
| $\tau_{(7–12).4}$ | 1.282/bimodal | 0.011 | 0.009 | 0.125 | 0.124 | −0.018 | 0.009 | 0.957 |
| $\alpha_{(1–6)}$ | 2.269/normal | 0.049 | 0.021 | 0.271 | 0.267 | −0.054 | 0.033 | 0.942 |
| $\alpha_{(7–12)}$ | 2.269/bimodal | 0.053 | 0.023 | 0.274 | 0.269 | −0.063 | 0.035 | 0.943 |
| $\beta_{(1–6).1}$ | −2.056/normal | −0.022 | 0.011 | 0.235 | 0.234 | | | |
| $\beta_{(1–6).2}$ | −0.804/normal | −0.003 | 0.004 | 0.128 | 0.128 | | | |
| $\beta_{(1–6).3}$ | 0.804/normal | 0.006 | 0.007 | 0.129 | 0.129 | | | |
| $\beta_{(1–6).4}$ | 2.056/normal | 0.024 | 0.012 | 0.237 | 0.236 | | | |
| $\beta_{(7–12).1}$ | −1.602/bimodal | −0.011 | 0.007 | 0.184 | 0.183 | | | |
| $\beta_{(7–12).2}$ | −0.157/bimodal | 0.001 | −0.008 | 0.112 | 0.112 | | | |
| $\beta_{(7–12).3}$ | 0.157/bimodal | 0.002 | 0.014 | 0.112 | 0.112 | | | |
| $\beta_{(7–12).4}$ | 1.602/bimodal | 0.012 | 0.008 | 0.184 | 0.183 | | | |

*Table E.27.* IRT-grm standardized parameters of interest and corresponding standard errors in Cell bnNS2 averaged over items with equal population value/shape. $n = 200$; $R = 1000$.

| | $\omega$ | PB($\hat{\omega}$) | RB($\hat{\omega}$) | RMSE($\hat{\omega}$) | SD($\hat{\omega}$) | RB($\hat{se}$) | RMSE($\hat{se}$) | 95%-cov. |
|---|---|---|---|---|---|---|---|---|
| $\lambda_{(1–6)}$ | 0.800/normal | −0.004 | −0.005 | 0.034 | 0.034 | −0.010 | 0.004 | 0.945 |
| $\lambda_{(7–12)}$ | 0.800/bimodal | −0.007 | −0.009 | 0.035 | 0.035 | −0.018 | 0.004 | 0.946 |
| $\tau_{(1–6).1}$ | −1.645/normal | −0.011 | 0.007 | 0.162 | 0.161 | −0.040 | 0.022 | 0.952 |
| $\tau_{(1–6).2}$ | −0.643/normal | 0.010 | −0.015 | 0.094 | 0.093 | −0.001 | 0.003 | 0.949 |
| $\tau_{(1–6).3}$ | 0.643/normal | −0.008 | −0.012 | 0.095 | 0.095 | −0.016 | 0.004 | 0.944 |
| $\tau_{(1–6).4}$ | 1.645/normal | 0.014 | 0.009 | 0.162 | 0.161 | −0.038 | 0.022 | 0.951 |
| $\tau_{(7–12).1}$ | −1.282/bimodal | 0.002 | −0.001 | 0.124 | 0.124 | −0.019 | 0.010 | 0.950 |
| $\tau_{(7–12).2}$ | −0.126/bimodal | 0.005 | −0.040 | 0.087 | 0.087 | −0.020 | 0.002 | 0.943 |
| $\tau_{(7–12).3}$ | 0.126/bimodal | −0.002 | −0.016 | 0.087 | 0.087 | −0.018 | 0.002 | 0.947 |
| $\tau_{(7–12).4}$ | 1.282/bimodal | 0.000 | 0.000 | 0.125 | 0.125 | −0.027 | 0.010 | 0.950 |
| $\alpha_{(1–6)}$ | 2.269/normal | −0.002 | −0.001 | 0.266 | 0.266 | −0.012 | 0.033 | 0.950 |
| $\alpha_{(7–12)}$ | 2.269/bimodal | −0.025 | −0.011 | 0.269 | 0.268 | −0.027 | 0.033 | 0.936 |
| $\beta_{(1–6).1}$ | −2.056/normal | −0.031 | 0.015 | 0.243 | 0.241 | | | |
| $\beta_{(1–6).2}$ | −0.804/normal | 0.006 | −0.008 | 0.126 | 0.126 | | | |
| $\beta_{(1–6).3}$ | 0.804/normal | −0.004 | −0.005 | 0.127 | 0.127 | | | |
| $\beta_{(1–6).4}$ | 2.056/normal | 0.034 | 0.017 | 0.245 | 0.243 | | | |
| $\beta_{(7–12).1}$ | −1.602/bimodal | −0.016 | 0.010 | 0.187 | 0.186 | | | |
| $\beta_{(7–12).2}$ | −0.157/bimodal | 0.004 | −0.028 | 0.111 | 0.111 | | | |
| $\beta_{(7–12).3}$ | 0.157/bimodal | −0.001 | −0.004 | 0.110 | 0.110 | | | |
| $\beta_{(7–12).4}$ | 1.602/bimodal | 0.019 | 0.012 | 0.188 | 0.187 | | | |

*Table E.28.* FA-lin standardized parameters of interest and corresponding standard errors in Cell bnNS6 averaged over items with equal population value/shape. $n = 600$; $R = 1000$.

| | $\omega$ | PB($\hat{\omega}$) | RB($\hat{\omega}$) | RMSE($\hat{\omega}$) | SD($\hat{\omega}$) | RB($\hat{se}$) | RMSE($\hat{se}$) | 95%-cov. |
|---|---|---|---|---|---|---|---|---|
| $\lambda_{(1–6)}$ | 0.800/normal | −0.046 | −0.057 | 0.050 | 0.019 | 0.010 | 0.001 | 0.306 |
| $\lambda_{(7–12)}$ | 0.800/bimodal | −0.049 | −0.061 | 0.052 | 0.019 | 0.014 | 0.001 | 0.259 |

*Table E.29.* FA-poly standardized parameters of interest and corresponding standard errors in Cell bnNS6 averaged over items with equal population value/shape. $n = 600$; $R = 1000$.

| | $\omega$ | PB($\hat{\omega}$) | RB($\hat{\omega}$) | RMSE($\hat{\omega}$) | SD($\hat{\omega}$) | RB($\hat{se}$) | RMSE($\hat{se}$) | 95%-cov. |
|---|---|---|---|---|---|---|---|---|
| $\lambda_{(1–6)}$ | 0.800/normal | 0.001 | 0.001 | 0.019 | 0.019 | −0.016 | 0.001 | 0.942 |
| $\lambda_{(7–12)}$ | 0.800/bimodal | 0.001 | 0.002 | 0.019 | 0.019 | −0.017 | 0.001 | 0.941 |
| $\tau_{(1–6).1}$ | −1.645/normal | −0.002 | 0.001 | 0.088 | 0.088 | −0.011 | 0.006 | 0.951 |
| $\tau_{(1–6).2}$ | −0.643/normal | −0.000 | 0.000 | 0.054 | 0.054 | 0.021 | 0.002 | 0.953 |
| $\tau_{(1–6).3}$ | 0.643/normal | 0.001 | 0.001 | 0.055 | 0.055 | 0.005 | 0.001 | 0.951 |
| $\tau_{(1–6).4}$ | 1.645/normal | 0.006 | 0.004 | 0.087 | 0.087 | 0.004 | 0.005 | 0.956 |
| $\tau_{(7–12).1}$ | −1.282/bimodal | −0.002 | 0.001 | 0.069 | 0.069 | 0.007 | 0.003 | 0.956 |
| $\tau_{(7–12).2}$ | −0.126/bimodal | −0.000 | 0.002 | 0.051 | 0.051 | 0.006 | 0.001 | 0.949 |
| $\tau_{(7–12).3}$ | 0.126/bimodal | −0.000 | −0.001 | 0.051 | 0.051 | 0.002 | 0.001 | 0.945 |
| $\tau_{(7–12).4}$ | 1.282/bimodal | 0.003 | 0.002 | 0.071 | 0.071 | −0.012 | 0.003 | 0.951 |
| $\alpha_{(1–6)}$ | 2.269/normal | 0.018 | 0.008 | 0.151 | 0.150 | −0.019 | 0.010 | 0.949 |
| $\alpha_{(7–12)}$ | 2.269/bimodal | 0.021 | 0.009 | 0.150 | 0.149 | −0.016 | 0.011 | 0.948 |
| $\beta_{(1–6).1}$ | −2.056/normal | −0.002 | 0.001 | 0.131 | 0.131 | | | |
| $\beta_{(1–6).2}$ | −0.804/normal | 0.000 | −0.000 | 0.072 | 0.072 | | | |
| $\beta_{(1–6).3}$ | 0.804/normal | 0.000 | 0.000 | 0.074 | 0.074 | | | |
| $\beta_{(1–6).4}$ | 2.056/normal | 0.006 | 0.003 | 0.129 | 0.129 | | | |
| $\beta_{(7–12).1}$ | −1.602/bimodal | −0.000 | 0.000 | 0.103 | 0.103 | | | |
| $\beta_{(7–12).2}$ | −0.157/bimodal | −0.000 | 0.001 | 0.064 | 0.064 | | | |
| $\beta_{(7–12).3}$ | 0.157/bimodal | −0.000 | −0.002 | 0.064 | 0.064 | | | |
| $\beta_{(7–12).4}$ | 1.602/bimodal | 0.002 | 0.001 | 0.105 | 0.105 | | | |

*Table E.30.* IRT-grm standardized parameters of interest and corresponding standard errors in Cell bnNS6 averaged over items with equal population value/shape. $n = 600$; $R = 1000$.

| | $\omega$ | PB($\hat{\omega}$) | RB($\hat{\omega}$) | RMSE($\hat{\omega}$) | SD($\hat{\omega}$) | RB($\hat{se}$) | RMSE($\hat{se}$) | 95%-cov. |
|---|---|---|---|---|---|---|---|---|
| $\lambda_{(1–6)}$ | 0.800/normal | −0.004 | −0.005 | 0.020 | 0.019 | 0.002 | 0.001 | 0.952 |
| $\lambda_{(7–12)}$ | 0.800/bimodal | −0.008 | −0.010 | 0.021 | 0.020 | −0.002 | 0.001 | 0.944 |
| $\tau_{(1–6).1}$ | −1.645/normal | 0.003 | −0.002 | 0.089 | 0.089 | −0.011 | 0.007 | 0.948 |
| $\tau_{(1–6).2}$ | −0.643/normal | 0.012 | −0.019 | 0.054 | 0.052 | 0.026 | 0.002 | 0.947 |
| $\tau_{(1–6).3}$ | 0.643/normal | −0.011 | −0.018 | 0.054 | 0.053 | 0.015 | 0.001 | 0.948 |
| $\tau_{(1–6).4}$ | 1.645/normal | −0.002 | −0.001 | 0.087 | 0.087 | 0.006 | 0.006 | 0.952 |
| $\tau_{(7–12).1}$ | −1.282/bimodal | 0.009 | −0.007 | 0.070 | 0.069 | 0.008 | 0.003 | 0.949 |
| $\tau_{(7–12).2}$ | −0.126/bimodal | 0.004 | −0.034 | 0.049 | 0.049 | 0.011 | 0.001 | 0.952 |
| $\tau_{(7–12).3}$ | 0.126/bimodal | −0.003 | −0.026 | 0.049 | 0.049 | 0.003 | 0.001 | 0.947 |
| $\tau_{(7–12).4}$ | 1.282/bimodal | −0.009 | −0.007 | 0.071 | 0.070 | −0.010 | 0.003 | 0.944 |
| $\alpha_{(1–6)}$ | 2.269/normal | −0.024 | −0.011 | 0.151 | 0.149 | −0.001 | 0.010 | 0.941 |
| $\alpha_{(7–12)}$ | 2.269/bimodal | −0.049 | −0.022 | 0.156 | 0.148 | −0.001 | 0.011 | 0.925 |
| $\beta_{(1–6).1}$ | −2.056/normal | −0.009 | 0.005 | 0.132 | 0.132 | | | |
| $\beta_{(1–6).2}$ | −0.804/normal | 0.011 | −0.013 | 0.071 | 0.070 | | | |
| $\beta_{(1–6).3}$ | 0.804/normal | −0.009 | −0.012 | 0.072 | 0.071 | | | |
| $\beta_{(1–6).4}$ | 2.056/normal | 0.011 | 0.005 | 0.130 | 0.130 | | | |
| $\beta_{(7–12).1}$ | −1.602/bimodal | −0.006 | 0.004 | 0.103 | 0.103 | | | |
| $\beta_{(7–12).2}$ | −0.157/bimodal | 0.004 | −0.024 | 0.062 | 0.062 | | | |
| $\beta_{(7–12).3}$ | 0.157/bimodal | −0.002 | −0.016 | 0.062 | 0.062 | | | |
| $\beta_{(7–12).4}$ | 1.602/bimodal | 0.005 | 0.003 | 0.105 | 0.105 | | | |

*Table E.31.* FA-lin standardized parameters of interest and corresponding standard errors in Cell nRS2 averaged over items with equal population value/shape. $n = 200$; $R = 1000$.

| | $\omega$ | PB($\hat{\omega}$) | RB($\hat{\omega}$) | RMSE($\hat{\omega}$) | SD($\hat{\omega}$) | RB($\hat{se}$) | RMSE($\hat{se}$) | 95%-cov. |
|---|---|---|---|---|---|---|---|---|
| $\lambda_{(1–12)}$ | 0.800/normal | −0.065 | −0.081 | 0.074 | 0.035 | 0.020 | 0.004 | 0.578 |

*Table E.32.* FA-poly standardized parameters of interest and corresponding standard errors in Cell nRS2 averaged over items with equal population value/shape. $n = 200$; $R = 1000$.

| | $\omega$ | PB($\hat{\omega}$) | RB($\hat{\omega}$) | RMSE($\hat{\omega}$) | SD($\hat{\omega}$) | RB($\hat{se}$) | RMSE($\hat{se}$) | 95%-cov. |
|---|---|---|---|---|---|---|---|---|
| $\lambda_{(1–12)}$ | 0.800/normal | −0.018 | −0.022 | 0.038 | 0.034 | −0.072 | 0.005 | 0.921 |
| $\tau_{(1–12).1}$ | −1.483/normal | −0.180 | 0.121 | 0.240 | 0.159 | −0.039 | 0.020 | 0.876 |
| $\tau_{(1–12).2}$ | −0.680/normal | 0.036 | −0.054 | 0.102 | 0.095 | 0.008 | 0.003 | 0.924 |
| $\tau_{(1–12).3}$ | 0.583/normal | 0.066 | 0.113 | 0.115 | 0.094 | 0.020 | 0.003 | 0.913 |
| $\tau_{(1–12).4}$ | 1.790/normal | −0.120 | −0.067 | 0.197 | 0.156 | −0.014 | 0.019 | 0.806 |
| $\alpha_{(1–12)}$ | 2.269/normal | −0.107 | −0.047 | 0.264 | 0.241 | −0.074 | 0.032 | 0.878 |
| $\beta_{(1–12).1}$ | −1.853/normal | −0.276 | 0.149 | 0.362 | 0.233 | | | |
| $\beta_{(1–12).2}$ | −0.850/normal | 0.026 | −0.030 | 0.131 | 0.128 | | | |
| $\beta_{(1–12).3}$ | 0.728/normal | 0.104 | 0.142 | 0.170 | 0.135 | | | |
| $\beta_{(1–12).4}$ | 2.237/normal | −0.096 | −0.043 | 0.268 | 0.250 | | | |

*Table E.33.* IRT-grm standardized parameters of interest and corresponding standard errors in Cell nRS2 averaged over items with equal population value/shape. $n = 200$; $R = 1000$.

| | $\omega$ | PB($\hat{\omega}$) | RB($\hat{\omega}$) | RMSE($\hat{\omega}$) | SD($\hat{\omega}$) | RB($\hat{se}$) | RMSE($\hat{se}$) | 95%-cov. |
|---|---|---|---|---|---|---|---|---|
| $\lambda_{(1–12)}$ | 0.800/normal | −0.009 | −0.012 | 0.037 | 0.036 | −0.005 | 0.005 | 0.949 |
| $\tau_{(1–12).1}$ | −1.483/normal | −0.032 | 0.021 | 0.144 | 0.140 | −0.034 | 0.022 | 0.962 |
| $\tau_{(1–12).2}$ | −0.680/normal | 0.012 | −0.017 | 0.082 | 0.081 | 0.030 | 0.004 | 0.955 |
| $\tau_{(1–12).3}$ | 0.583/normal | 0.010 | 0.017 | 0.098 | 0.098 | 0.041 | 0.006 | 0.957 |
| $\tau_{(1–12).4}$ | 1.790/normal | 0.007 | 0.004 | 0.179 | 0.179 | −0.010 | 0.023 | 0.955 |
| $\alpha_{(1–12)}$ | 2.269/normal | −0.041 | −0.018 | 0.275 | 0.272 | −0.011 | 0.035 | 0.932 |
| $\beta_{(1–12).1}$ | −1.853/normal | −0.068 | 0.037 | 0.229 | 0.219 | | | |
| $\beta_{(1–12).2}$ | −0.850/normal | 0.003 | −0.004 | 0.108 | 0.108 | | | |
| $\beta_{(1–12).3}$ | 0.728/normal | 0.024 | 0.034 | 0.139 | 0.137 | | | |
| $\beta_{(1–12).4}$ | 2.237/normal | 0.043 | 0.019 | 0.276 | 0.273 | | | |

*Table E.34.* FA-lin standardized parameters of interest and corresponding standard errors in Cell nRS6 averaged over items with equal population value/shape. $n = 600$; $R = 1000$.

| | $\omega$ | PB($\hat{\omega}$) | RB($\hat{\omega}$) | RMSE($\hat{\omega}$) | SD($\hat{\omega}$) | RB($\hat{se}$) | RMSE($\hat{se}$) | 95%-cov. |
|---|---|---|---|---|---|---|---|---|
| $\lambda_{(1–12)}$ | 0.800/normal | −0.064 | −0.080 | 0.067 | 0.020 | 0.009 | 0.001 | 0.081 |

*Table E.35.* FA-poly standardized parameters of interest and corresponding standard errors in Cell nRS6 averaged over items with equal population value/shape. $n = 600$; $R = 1000$.

| | $\omega$ | PB($\hat{\omega}$) | RB($\hat{\omega}$) | RMSE($\hat{\omega}$) | SD($\hat{\omega}$) | RB($\hat{se}$) | RMSE($\hat{se}$) | 95%-cov. |
|---|---|---|---|---|---|---|---|---|
| $\lambda_{(1–12)}$ | 0.800/normal | −0.019 | −0.024 | 0.027 | 0.019 | −0.044 | 0.002 | 0.845 |
| $\tau_{(1–12).1}$ | −1.483/normal | −0.170 | 0.115 | 0.191 | 0.088 | −0.010 | 0.006 | 0.516 |
| $\tau_{(1–12).2}$ | −0.680/normal | 0.034 | −0.050 | 0.065 | 0.055 | 0.004 | 0.001 | 0.896 |
| $\tau_{(1–12).3}$ | 0.583/normal | 0.061 | 0.104 | 0.083 | 0.057 | −0.024 | 0.002 | 0.807 |
| $\tau_{(1–12).4}$ | 1.790/normal | −0.141 | −0.079 | 0.167 | 0.090 | −0.032 | 0.006 | 0.610 |
| $\alpha_{(1–12)}$ | 2.269/normal | −0.132 | −0.058 | 0.190 | 0.137 | −0.045 | 0.011 | 0.790 |
| $\beta_{(1–12).1}$ | −1.853/normal | −0.264 | 0.142 | 0.294 | 0.130 | | | |
| $\beta_{(1–12).2}$ | −0.850/normal | 0.022 | −0.026 | 0.077 | 0.074 | | | |
| $\beta_{(1–12).3}$ | 0.728/normal | 0.096 | 0.132 | 0.126 | 0.081 | | | |
| $\beta_{(1–12).4}$ | 2.237/normal | −0.124 | −0.055 | 0.190 | 0.144 | | | |

*Table E.36.* IRT-grm standardized parameters of interest and corresponding standard errors in Cell nRS6 averaged over items with equal population value/shape. $n = 600$; $R = 1000$.

| | $\omega$ | PB($\hat{\omega}$) | RB($\hat{\omega}$) | RMSE($\hat{\omega}$) | SD($\hat{\omega}$) | RB($\hat{se}$) | RMSE($\hat{se}$) | 95%-cov. |
|---|---|---|---|---|---|---|---|---|
| $\lambda_{(1–12)}$ | 0.800/normal | −0.010 | −0.012 | 0.023 | 0.021 | −0.011 | 0.002 | 0.935 |
| $\tau_{(1–12).1}$ | −1.483/normal | −0.023 | 0.016 | 0.081 | 0.078 | −0.015 | 0.006 | 0.951 |
| $\tau_{(1–12).2}$ | −0.680/normal | 0.010 | −0.015 | 0.049 | 0.048 | 0.006 | 0.001 | 0.945 |
| $\tau_{(1–12).3}$ | 0.583/normal | 0.006 | 0.010 | 0.060 | 0.059 | −0.016 | 0.002 | 0.944 |
| $\tau_{(1–12).4}$ | 1.790/normal | −0.018 | −0.010 | 0.104 | 0.102 | −0.021 | 0.007 | 0.938 |
| $\alpha_{(1–12)}$ | 2.269/normal | −0.066 | −0.029 | 0.168 | 0.154 | −0.014 | 0.012 | 0.910 |
| $\beta_{(1–12).1}$ | −1.853/normal | −0.055 | 0.030 | 0.134 | 0.122 | | | |
| $\beta_{(1–12).2}$ | −0.850/normal | 0.002 | −0.002 | 0.063 | 0.063 | | | |
| $\beta_{(1–12).3}$ | 0.728/normal | 0.017 | 0.024 | 0.085 | 0.083 | | | |
| $\beta_{(1–12).4}$ | 2.237/normal | 0.007 | 0.003 | 0.156 | 0.156 | | | |

*Table E.37.* FA-lin standardized parameters of interest and corresponding standard errors in Cell rnRS2 averaged over items with equal population value/shape. $n = 200$; $R = 1000$.

| | $\omega$ | PB($\hat{\omega}$) | RB($\hat{\omega}$) | RMSE($\hat{\omega}$) | SD($\hat{\omega}$) | RB($\hat{se}$) | RMSE($\hat{se}$) | 95%-cov. |
|---|---|---|---|---|---|---|---|---|
| $\lambda_{(1–6)}$ | 0.800/normal | −0.081 | −0.102 | 0.088 | 0.034 | 0.077 | 0.004 | 0.374 |
| $\lambda_{(7–12)}$ | 0.800/right-skewed | −0.016 | −0.021 | 0.042 | 0.038 | −0.226 | 0.010 | 0.870 |

*Table E.38.* FA-poly standardized parameters of interest and corresponding standard errors in Cell rnRS2 averaged over items with equal population value/shape. $n = 200$; $R = 1000$.

| | $\omega$ | PB($\hat{\omega}$) | RB($\hat{\omega}$) | RMSE($\hat{\omega}$) | SD($\hat{\omega}$) | RB($\hat{se}$) | RMSE($\hat{se}$) | 95%-cov. |
|---|---|---|---|---|---|---|---|---|
| $\lambda_{(1–6)}$ | 0.800/normal | −0.009 | −0.011 | 0.034 | 0.033 | −0.126 | 0.006 | 0.915 |
| $\lambda_{(7–12)}$ | 0.800/right-skewed | 0.058 | 0.073 | 0.069 | 0.036 | −0.154 | 0.008 | 0.493 |
| $\tau_{(1–6).1}$ | −1.483/normal | −0.180 | 0.121 | 0.240 | 0.159 | −0.037 | 0.020 | 0.880 |
| $\tau_{(1–6).2}$ | −0.680/normal | 0.033 | −0.049 | 0.102 | 0.097 | −0.008 | 0.003 | 0.927 |
| $\tau_{(1–6).3}$ | 0.583/normal | 0.066 | 0.113 | 0.118 | 0.098 | −0.025 | 0.004 | 0.904 |
| $\tau_{(1–6).4}$ | 1.790/normal | −0.117 | −0.066 | 0.200 | 0.161 | −0.044 | 0.021 | 0.801 |
| $\tau_{(7–12).1}$ | 0.260/right-skewed | 0.088 | 0.337 | 0.128 | 0.094 | −0.034 | 0.004 | 0.841 |
| $\tau_{(7–12).2}$ | 0.760/right-skewed | 0.046 | 0.060 | 0.113 | 0.104 | −0.035 | 0.005 | 0.923 |
| $\tau_{(7–12).3}$ | 1.260/right-skewed | −0.026 | −0.021 | 0.125 | 0.123 | −0.031 | 0.008 | 0.934 |
| $\tau_{(7–12).4}$ | 1.760/right-skewed | −0.115 | −0.065 | 0.194 | 0.157 | −0.033 | 0.018 | 0.839 |
| $\alpha_{(1–6)}$ | 2.269/normal | −0.044 | −0.020 | 0.250 | 0.246 | −0.124 | 0.039 | 0.897 |
| $\alpha_{(7–12)}$ | 2.269/right-skewed | 0.664 | 0.292 | 0.835 | 0.506 | −0.176 | 0.150 | 0.705 |
| $\beta_{(1–6).1}$ | −1.853/normal | −0.253 | 0.136 | 0.339 | 0.226 | | | |
| $\beta_{(1–6).2}$ | −0.850/normal | 0.031 | −0.036 | 0.132 | 0.128 | | | |
| $\beta_{(1–6).3}$ | 0.728/normal | 0.095 | 0.130 | 0.168 | 0.139 | | | |
| $\beta_{(1–6).4}$ | 2.237/normal | −0.116 | −0.052 | 0.280 | 0.255 | | | |
| $\beta_{(7–12).1}$ | 0.325/right-skewed | 0.080 | 0.248 | 0.136 | 0.110 | | | |
| $\beta_{(7–12).2}$ | 0.950/right-skewed | −0.009 | −0.009 | 0.135 | 0.135 | | | |
| $\beta_{(7–12).3}$ | 1.575/right-skewed | −0.133 | −0.084 | 0.219 | 0.174 | | | |
| $\beta_{(7–12).4}$ | 2.200/right-skewed | −0.276 | −0.126 | 0.358 | 0.229 | | | |

*Table E.39.* IRT-grm standardized parameters of interest and corresponding standard errors in Cell rnRS2 averaged over items with equal population value/shape. $n = 200$; $R = 1000$.

| | $\omega$ | PB($\hat{\omega}$) | RB($\hat{\omega}$) | RMSE($\hat{\omega}$) | SD($\hat{\omega}$) | RB($\hat{se}$) | RMSE($\hat{se}$) | 95%-cov. |
|---|---|---|---|---|---|---|---|---|
| $\lambda_{(1–6)}$ | 0.800/normal | −0.012 | −0.015 | 0.038 | 0.036 | 0.007 | 0.005 | 0.952 |
| $\lambda_{(7–12)}$ | 0.800/right-skewed | 0.006 | 0.008 | 0.037 | 0.037 | −0.010 | 0.006 | 0.923 |
| $\tau_{(1–6).1}$ | −1.483/normal | −0.058 | 0.039 | 0.156 | 0.145 | −0.036 | 0.022 | 0.957 |
| $\tau_{(1–6).2}$ | −0.680/normal | 0.015 | −0.022 | 0.086 | 0.085 | 0.004 | 0.003 | 0.949 |
| $\tau_{(1–6).3}$ | 0.583/normal | 0.013 | 0.022 | 0.101 | 0.100 | 0.003 | 0.005 | 0.952 |
| $\tau_{(1–6).4}$ | 1.790/normal | −0.009 | −0.005 | 0.181 | 0.181 | −0.036 | 0.024 | 0.944 |
| $\tau_{(7–12).1}$ | 0.260/right-skewed | 0.019 | 0.071 | 0.094 | 0.092 | −0.008 | 0.003 | 0.946 |
| $\tau_{(7–12).2}$ | 0.760/right-skewed | 0.002 | 0.003 | 0.108 | 0.108 | −0.015 | 0.006 | 0.948 |
| $\tau_{(7–12).3}$ | 1.260/right-skewed | −0.010 | −0.008 | 0.134 | 0.134 | −0.016 | 0.011 | 0.947 |
| $\tau_{(7–12).4}$ | 1.760/right-skewed | −0.011 | −0.006 | 0.175 | 0.175 | −0.022 | 0.021 | 0.947 |
| $\alpha_{(1–6)}$ | 2.269/normal | −0.062 | −0.027 | 0.274 | 0.267 | 0.005 | 0.035 | 0.932 |
| $\alpha_{(7–12)}$ | 2.269/right-skewed | 0.091 | 0.040 | 0.329 | 0.316 | −0.020 | 0.047 | 0.953 |
| $\beta_{(1–6).1}$ | −1.853/normal | −0.108 | 0.058 | 0.246 | 0.221 | | | |
| $\beta_{(1–6).2}$ | −0.850/normal | 0.004 | −0.005 | 0.113 | 0.113 | | | |
| $\beta_{(1–6).3}$ | 0.728/normal | 0.030 | 0.042 | 0.144 | 0.141 | | | |
| $\beta_{(1–6).4}$ | 2.237/normal | 0.030 | 0.014 | 0.281 | 0.280 | | | |
| $\beta_{(7–12).1}$ | 0.325/right-skewed | 0.022 | 0.067 | 0.120 | 0.118 | | | |
| $\beta_{(7–12).2}$ | 0.950/right-skewed | −0.001 | −0.001 | 0.151 | 0.151 | | | |
| $\beta_{(7–12).3}$ | 1.575/right-skewed | −0.020 | −0.013 | 0.200 | 0.199 | | | |
| $\beta_{(7–12).4}$ | 2.200/right-skewed | −0.023 | −0.011 | 0.267 | 0.266 | | | |

*Table E.40.* FA-lin standardized parameters of interest and corresponding standard errors in Cell rnRS6 averaged over items with equal population value/shape. $n = 600$; $R = 1000$.

| | $\omega$ | PB($\hat{\omega}$) | RB($\hat{\omega}$) | RMSE($\hat{\omega}$) | SD($\hat{\omega}$) | RB($\hat{se}$) | RMSE($\hat{se}$) | 95%-cov. |
|---|---|---|---|---|---|---|---|---|
| $\lambda_{(1–6)}$ | 0.800/normal | −0.080 | −0.100 | 0.082 | 0.020 | 0.066 | 0.002 | 0.008 |
| $\lambda_{(7–12)}$ | 0.800/right-skewed | −0.014 | −0.018 | 0.026 | 0.022 | −0.211 | 0.005 | 0.822 |

*Table E.41.* FA-poly standardized parameters of interest and corresponding standard errors in Cell rnRS6 averaged over items with equal population value/shape. $n = 600$; $R = 1000$.

| | $\omega$ | PB($\hat{\omega}$) | RB($\hat{\omega}$) | RMSE($\hat{\omega}$) | SD($\hat{\omega}$) | RB($\hat{se}$) | RMSE($\hat{se}$) | 95%-cov. |
|---|---|---|---|---|---|---|---|---|
| $\lambda_{(1–6)}$ | 0.800/normal | −0.011 | −0.014 | 0.022 | 0.019 | −0.111 | 0.003 | 0.889 |
| $\lambda_{(7–12)}$ | 0.800/right-skewed | 0.058 | 0.073 | 0.062 | 0.021 | −0.120 | 0.003 | 0.178 |
| $\tau_{(1–6).1}$ | −1.483/normal | −0.169 | 0.114 | 0.190 | 0.088 | −0.015 | 0.006 | 0.522 |
| $\tau_{(1–6).2}$ | −0.680/normal | 0.036 | −0.053 | 0.065 | 0.054 | 0.028 | 0.002 | 0.899 |
| $\tau_{(1–6).3}$ | 0.583/normal | 0.062 | 0.106 | 0.082 | 0.054 | 0.024 | 0.002 | 0.810 |
| $\tau_{(1–6).4}$ | 1.790/normal | −0.138 | −0.077 | 0.164 | 0.088 | −0.007 | 0.006 | 0.623 |
| $\tau_{(7–12).1}$ | 0.260/right-skewed | 0.086 | 0.332 | 0.100 | 0.051 | 0.019 | 0.001 | 0.610 |
| $\tau_{(7–12).2}$ | 0.760/right-skewed | 0.041 | 0.054 | 0.071 | 0.057 | 0.009 | 0.002 | 0.895 |
| $\tau_{(7–12).3}$ | 1.260/right-skewed | −0.036 | −0.029 | 0.076 | 0.066 | 0.024 | 0.003 | 0.919 |
| $\tau_{(7–12).4}$ | 1.760/right-skewed | −0.132 | −0.075 | 0.158 | 0.087 | −0.014 | 0.005 | 0.654 |
| $\alpha_{(1–6)}$ | 2.269/normal | −0.076 | −0.033 | 0.159 | 0.140 | −0.112 | 0.018 | 0.860 |
| $\alpha_{(7–12)}$ | 2.269/right-skewed | 0.602 | 0.265 | 0.660 | 0.270 | −0.126 | 0.047 | 0.263 |
| $\beta_{(1–6).1}$ | −1.853/normal | −0.241 | 0.130 | 0.272 | 0.126 | | | |
| $\beta_{(1–6).2}$ | −0.850/normal | 0.033 | −0.039 | 0.079 | 0.072 | | | |
| $\beta_{(1–6).3}$ | 0.728/normal | 0.090 | 0.123 | 0.118 | 0.076 | | | |
| $\beta_{(1–6).4}$ | 2.237/normal | −0.141 | −0.063 | 0.198 | 0.139 | | | |
| $\beta_{(7–12).1}$ | 0.325/right-skewed | 0.079 | 0.242 | 0.099 | 0.060 | | | |
| $\beta_{(7–12).2}$ | 0.950/right-skewed | −0.015 | −0.016 | 0.075 | 0.074 | | | |
| $\beta_{(7–12).3}$ | 1.575/right-skewed | −0.147 | −0.094 | 0.175 | 0.094 | | | |
| $\beta_{(7–12).4}$ | 2.200/right-skewed | −0.301 | −0.137 | 0.327 | 0.126 | | | |

*Table E.42.* IRT-grm standardized parameters of interest and corresponding standard errors in Cell rnRS6 averaged over items with equal population value/shape. $n = 600$; $R = 1000$.

| | $\omega$ | PB($\hat{\omega}$) | RB($\hat{\omega}$) | RMSE($\hat{\omega}$) | SD($\hat{\omega}$) | RB($\hat{se}$) | RMSE($\hat{se}$) | 95%-cov. |
|---|---|---|---|---|---|---|---|---|
| $\lambda_{(1–6)}$ | 0.800/normal | −0.013 | −0.016 | 0.024 | 0.021 | −0.007 | 0.002 | 0.921 |
| $\lambda_{(7–12)}$ | 0.800/right-skewed | 0.006 | 0.008 | 0.022 | 0.021 | 0.014 | 0.002 | 0.930 |
| $\tau_{(1–6).1}$ | −1.483/normal | −0.045 | 0.030 | 0.092 | 0.080 | −0.017 | 0.006 | 0.929 |
| $\tau_{(1–6).2}$ | −0.680/normal | 0.018 | −0.027 | 0.052 | 0.048 | 0.020 | 0.002 | 0.937 |
| $\tau_{(1–6).3}$ | 0.583/normal | 0.010 | 0.016 | 0.056 | 0.056 | 0.036 | 0.002 | 0.958 |
| $\tau_{(1–6).4}$ | 1.790/normal | −0.031 | −0.018 | 0.103 | 0.098 | 0.007 | 0.007 | 0.934 |
| $\tau_{(7–12).1}$ | 0.260/right-skewed | 0.018 | 0.069 | 0.054 | 0.051 | 0.026 | 0.002 | 0.946 |
| $\tau_{(7–12).2}$ | 0.760/right-skewed | −0.000 | −0.000 | 0.060 | 0.060 | 0.020 | 0.002 | 0.959 |
| $\tau_{(7–12).3}$ | 1.260/right-skewed | −0.018 | −0.014 | 0.074 | 0.072 | 0.041 | 0.004 | 0.948 |
| $\tau_{(7–12).4}$ | 1.760/right-skewed | −0.029 | −0.016 | 0.101 | 0.097 | −0.004 | 0.006 | 0.931 |
| $\alpha_{(1–6)}$ | 2.269/normal | −0.086 | −0.038 | 0.176 | 0.153 | −0.008 | 0.011 | 0.889 |
| $\alpha_{(7–12)}$ | 2.269/right-skewed | 0.061 | 0.027 | 0.182 | 0.171 | 0.012 | 0.014 | 0.949 |
| $\beta_{(1–6).1}$ | −1.853/normal | −0.089 | 0.048 | 0.152 | 0.123 | | | |
| $\beta_{(1–6).2}$ | −0.850/normal | 0.009 | −0.011 | 0.065 | 0.065 | | | |
| $\beta_{(1–6).3}$ | 0.728/normal | 0.025 | 0.034 | 0.082 | 0.078 | | | |
| $\beta_{(1–6).4}$ | 2.237/normal | −0.001 | −0.001 | 0.151 | 0.151 | | | |
| $\beta_{(7–12).1}$ | 0.325/right-skewed | 0.020 | 0.062 | 0.068 | 0.065 | | | |
| $\beta_{(7–12).2}$ | 0.950/right-skewed | −0.006 | −0.007 | 0.083 | 0.083 | | | |
| $\beta_{(7–12).3}$ | 1.575/right-skewed | −0.033 | −0.021 | 0.112 | 0.107 | | | |
| $\beta_{(7–12).4}$ | 2.200/right-skewed | −0.050 | −0.023 | 0.154 | 0.145 | | | |

*Table E.43.* FA-lin standardized parameters of interest and corresponding standard errors in Cell lnRS2 averaged over items with equal population value/shape. $n = 200$; $R = 1000$.

| | $\omega$ | PB($\hat{\omega}$) | RB($\hat{\omega}$) | RMSE($\hat{\omega}$) | SD($\hat{\omega}$) | RB($\hat{se}$) | RMSE($\hat{se}$) | 95%-cov. |
|---|---|---|---|---|---|---|---|---|
| $\lambda_{(1–6)}$ | 0.800/normal | −0.073 | −0.091 | 0.082 | 0.038 | −0.013 | 0.004 | 0.526 |
| $\lambda_{(7–12)}$ | 0.800/left-skewed | −0.236 | −0.295 | 0.240 | 0.046 | 0.132 | 0.007 | 0.000 |

*Table E.44.* FA-poly standardized parameters of interest and corresponding standard errors in Cell lnRS2 averaged over items with equal population value/shape. $n = 200$; $R = 1000$.

| | $\omega$ | PB($\hat{\omega}$) | RB($\hat{\omega}$) | RMSE($\hat{\omega}$) | SD($\hat{\omega}$) | RB($\hat{se}$) | RMSE($\hat{se}$) | 95%-cov. |
|---|---|---|---|---|---|---|---|---|
| $\lambda_{(1–6)}$ | 0.800/normal | −0.019 | −0.024 | 0.041 | 0.036 | −0.021 | 0.004 | 0.933 |
| $\lambda_{(7–12)}$ | 0.800/left-skewed | −0.100 | −0.125 | 0.110 | 0.045 | 0.025 | 0.005 | 0.421 |
| $\tau_{(1–6).1}$ | −1.483/normal | −0.184 | 0.124 | 0.244 | 0.161 | −0.047 | 0.021 | 0.871 |
| $\tau_{(1–6).2}$ | −0.680/normal | 0.032 | −0.048 | 0.101 | 0.095 | 0.005 | 0.003 | 0.930 |
| $\tau_{(1–6).3}$ | 0.583/normal | 0.061 | 0.104 | 0.113 | 0.095 | 0.004 | 0.003 | 0.918 |
| $\tau_{(1–6).4}$ | 1.790/normal | −0.126 | −0.070 | 0.200 | 0.155 | −0.010 | 0.018 | 0.800 |
| $\tau_{(7–12).1}$ | −1.465/left-skewed | −0.175 | 0.120 | 0.235 | 0.157 | −0.040 | 0.018 | 0.822 |
| $\tau_{(7–12).2}$ | −1.156/left-skewed | −0.075 | 0.065 | 0.142 | 0.120 | −0.014 | 0.008 | 0.936 |
| $\tau_{(7–12).3}$ | −0.813/left-skewed | 0.008 | −0.009 | 0.102 | 0.102 | −0.017 | 0.004 | 0.950 |
| $\tau_{(7–12).4}$ | −0.416/left-skewed | 0.068 | −0.162 | 0.114 | 0.091 | −0.005 | 0.002 | 0.869 |
| $\alpha_{(1–6)}$ | 2.269/normal | −0.114 | −0.050 | 0.284 | 0.260 | −0.026 | 0.036 | 0.886 |
| $\alpha_{(7–12)}$ | 2.269/left-skewed | −0.583 | −0.257 | 0.622 | 0.216 | 0.024 | 0.027 | 0.271 |
| $\beta_{(1–6).1}$ | −1.853/normal | −0.287 | 0.155 | 0.374 | 0.240 | | | |
| $\beta_{(1–6).2}$ | −0.850/normal | 0.019 | −0.022 | 0.131 | 0.130 | | | |
| $\beta_{(1–6).3}$ | 0.728/normal | 0.099 | 0.136 | 0.171 | 0.139 | | | |
| $\beta_{(1–6).4}$ | 2.237/normal | −0.098 | −0.044 | 0.272 | 0.253 | | | |
| $\beta_{(7–12).1}$ | −1.831/left-skewed | −0.527 | 0.288 | 0.608 | 0.304 | | | |
| $\beta_{(7–12).2}$ | −1.445/left-skewed | −0.324 | 0.224 | 0.396 | 0.229 | | | |
| $\beta_{(7–12).3}$ | −1.016/left-skewed | −0.141 | 0.139 | 0.225 | 0.175 | | | |
| $\beta_{(7–12).4}$ | −0.521/left-skewed | 0.020 | −0.038 | 0.138 | 0.137 | | | |

*Table E.45.* IRT-grm standardized parameters of interest and corresponding standard errors in Cell lnRS2 averaged over items with equal population value/shape. $n = 200$; $R = 1000$.

| | $\omega$ | PB($\hat{\omega}$) | RB($\hat{\omega}$) | RMSE($\hat{\omega}$) | SD($\hat{\omega}$) | RB($\hat{se}$) | RMSE($\hat{se}$) | 95%-cov. |
|---|---|---|---|---|---|---|---|---|
| $\lambda_{(1–6)}$ | 0.800/normal | −0.011 | −0.013 | 0.038 | 0.036 | −0.004 | 0.005 | 0.952 |
| $\lambda_{(7–12)}$ | 0.800/left-skewed | −0.046 | −0.058 | 0.067 | 0.049 | 0.007 | 0.007 | 0.895 |
| $\tau_{(1–6).1}$ | −1.483/normal | −0.048 | 0.032 | 0.150 | 0.143 | −0.040 | 0.023 | 0.959 |
| $\tau_{(1–6).2}$ | −0.680/normal | 0.007 | −0.010 | 0.083 | 0.083 | 0.025 | 0.003 | 0.956 |
| $\tau_{(1–6).3}$ | 0.583/normal | 0.016 | 0.028 | 0.100 | 0.099 | 0.020 | 0.005 | 0.953 |
| $\tau_{(1–6).4}$ | 1.790/normal | −0.018 | −0.010 | 0.173 | 0.172 | −0.007 | 0.021 | 0.951 |
| $\tau_{(7–12).1}$ | −1.465/left-skewed | −0.060 | 0.041 | 0.157 | 0.145 | −0.030 | 0.023 | 0.961 |
| $\tau_{(7–12).2}$ | −1.156/left-skewed | −0.012 | 0.010 | 0.105 | 0.104 | 0.002 | 0.009 | 0.955 |
| $\tau_{(7–12).3}$ | −0.813/left-skewed | 0.012 | −0.015 | 0.088 | 0.087 | 0.009 | 0.004 | 0.951 |
| $\tau_{(7–12).4}$ | −0.416/left-skewed | 0.026 | −0.063 | 0.087 | 0.083 | 0.021 | 0.003 | 0.945 |
| $\alpha_{(1–6)}$ | 2.269/normal | −0.050 | −0.022 | 0.280 | 0.275 | −0.009 | 0.035 | 0.931 |
| $\alpha_{(7–12)}$ | 2.269/left-skewed | −0.277 | −0.122 | 0.415 | 0.309 | −0.007 | 0.050 | 0.790 |
| $\beta_{(1–6).1}$ | −1.853/normal | −0.092 | 0.050 | 0.239 | 0.221 | | | |
| $\beta_{(1–6).2}$ | −0.850/normal | −0.004 | 0.005 | 0.109 | 0.109 | | | |
| $\beta_{(1–6).3}$ | 0.728/normal | 0.034 | 0.046 | 0.143 | 0.139 | | | |
| $\beta_{(1–6).4}$ | 2.237/normal | 0.015 | 0.007 | 0.269 | 0.269 | | | |
| $\beta_{(7–12).1}$ | −1.831/left-skewed | −0.204 | 0.112 | 0.338 | 0.269 | | | |
| $\beta_{(7–12).2}$ | −1.445/left-skewed | −0.113 | 0.078 | 0.220 | 0.189 | | | |
| $\beta_{(7–12).3}$ | −1.016/left-skewed | −0.050 | 0.049 | 0.145 | 0.136 | | | |
| $\beta_{(7–12).4}$ | −0.521/left-skewed | 0.002 | −0.005 | 0.109 | 0.109 | | | |

*Table E.46.* FA-lin standardized parameters of interest and corresponding standard errors in Cell lnRS6 averaged over items with equal population value/shape. $n = 600$; $R = 1000$.

| | $\omega$ | PB($\hat{\omega}$) | RB($\hat{\omega}$) | RMSE($\hat{\omega}$) | SD($\hat{\omega}$) | RB($\hat{se}$) | RMSE($\hat{se}$) | 95%-cov. |
|---|---|---|---|---|---|---|---|---|
| $\lambda_{(1–6)}$ | 0.800/normal | −0.073 | −0.091 | 0.076 | 0.022 | −0.009 | 0.001 | 0.052 |
| $\lambda_{(7–12)}$ | 0.800/left-skewed | −0.237 | −0.296 | 0.239 | 0.026 | 0.156 | 0.004 | 0.000 |

*Table E.47.* FA-poly standardized parameters of interest and corresponding standard errors in Cell lnRS6 averaged over items with equal population value/shape. $n = 600$; $R = 1000$.

| | $\omega$ | PB($\hat{\omega}$) | RB($\hat{\omega}$) | RMSE($\hat{\omega}$) | SD($\hat{\omega}$) | RB($\hat{se}$) | RMSE($\hat{se}$) | 95%-cov. |
|---|---|---|---|---|---|---|---|---|
| $\lambda_{(1–6)}$ | 0.800/normal | −0.021 | −0.026 | 0.030 | 0.021 | 0.017 | 0.002 | 0.863 |
| $\lambda_{(7–12)}$ | 0.800/left-skewed | −0.104 | −0.130 | 0.107 | 0.026 | 0.079 | 0.003 | 0.013 |
| $\tau_{(1–6).1}$ | −1.483/normal | −0.167 | 0.113 | 0.188 | 0.087 | −0.000 | 0.006 | 0.532 |
| $\tau_{(1–6).2}$ | −0.680/normal | 0.034 | −0.050 | 0.065 | 0.056 | −0.013 | 0.002 | 0.893 |
| $\tau_{(1–6).3}$ | 0.583/normal | 0.060 | 0.102 | 0.081 | 0.055 | −0.003 | 0.001 | 0.816 |
| $\tau_{(1–6).4}$ | 1.790/normal | −0.138 | −0.077 | 0.163 | 0.087 | −0.006 | 0.005 | 0.623 |
| $\tau_{(7–12).1}$ | −1.465/left-skewed | −0.164 | 0.112 | 0.185 | 0.087 | −0.021 | 0.006 | 0.556 |
| $\tau_{(7–12).2}$ | −1.156/left-skewed | −0.070 | 0.061 | 0.098 | 0.068 | −0.006 | 0.002 | 0.847 |
| $\tau_{(7–12).3}$ | −0.813/left-skewed | 0.010 | −0.012 | 0.060 | 0.059 | −0.018 | 0.002 | 0.941 |
| $\tau_{(7–12).4}$ | −0.416/left-skewed | 0.069 | −0.165 | 0.087 | 0.053 | −0.016 | 0.001 | 0.718 |
| $\alpha_{(1–6)}$ | 2.269/normal | −0.147 | −0.065 | 0.207 | 0.146 | 0.017 | 0.012 | 0.805 |
| $\alpha_{(7–12)}$ | 2.269/left-skewed | −0.613 | −0.270 | 0.625 | 0.120 | 0.081 | 0.013 | 0.006 |
| $\beta_{(1–6).1}$ | −1.853/normal | −0.267 | 0.144 | 0.297 | 0.130 | | | |
| $\beta_{(1–6).2}$ | −0.850/normal | 0.020 | −0.023 | 0.079 | 0.076 | | | |
| $\beta_{(1–6).3}$ | 0.728/normal | 0.097 | 0.134 | 0.126 | 0.080 | | | |
| $\beta_{(1–6).4}$ | 2.237/normal | −0.114 | −0.051 | 0.182 | 0.142 | | | |
| $\beta_{(7–12).1}$ | −1.831/left-skewed | −0.513 | 0.280 | 0.541 | 0.170 | | | |
| $\beta_{(7–12).2}$ | −1.445/left-skewed | −0.320 | 0.222 | 0.346 | 0.130 | | | |
| $\beta_{(7–12).3}$ | −1.016/left-skewed | −0.139 | 0.137 | 0.171 | 0.100 | | | |
| $\beta_{(7–12).4}$ | −0.521/left-skewed | 0.020 | −0.039 | 0.081 | 0.079 | | | |

*Table E.48.* IRT-grm standardized parameters of interest and corresponding standard errors in Cell lnRS6 averaged over items with equal population value/shape. $n = 600$; $R = 1000$.

| | $\omega$ | PB($\hat{\omega}$) | RB($\hat{\omega}$) | RMSE($\hat{\omega}$) | SD($\hat{\omega}$) | RB($\hat{se}$) | RMSE($\hat{se}$) | 95%-cov. |
|---|---|---|---|---|---|---|---|---|
| $\lambda_{(1–6)}$ | 0.800/normal | −0.012 | −0.015 | 0.024 | 0.021 | −0.005 | 0.002 | 0.929 |
| $\lambda_{(7–12)}$ | 0.800/left-skewed | −0.047 | −0.058 | 0.054 | 0.028 | 0.015 | 0.002 | 0.663 |
| $\tau_{(1–6).1}$ | −1.483/normal | −0.034 | 0.023 | 0.084 | 0.077 | 0.002 | 0.006 | 0.947 |
| $\tau_{(1–6).2}$ | −0.680/normal | 0.008 | −0.012 | 0.050 | 0.049 | 0.001 | 0.001 | 0.950 |
| $\tau_{(1–6).3}$ | 0.583/normal | 0.015 | 0.026 | 0.059 | 0.057 | 0.012 | 0.002 | 0.947 |
| $\tau_{(1–6).4}$ | 1.790/normal | −0.033 | −0.018 | 0.102 | 0.097 | −0.002 | 0.006 | 0.927 |
| $\tau_{(7–12).1}$ | −1.465/left-skewed | −0.049 | 0.034 | 0.095 | 0.081 | −0.016 | 0.007 | 0.926 |
| $\tau_{(7–12).2}$ | −1.156/left-skewed | −0.008 | 0.007 | 0.060 | 0.060 | 0.002 | 0.003 | 0.951 |
| $\tau_{(7–12).3}$ | −0.813/left-skewed | 0.015 | −0.018 | 0.053 | 0.051 | −0.005 | 0.001 | 0.937 |
| $\tau_{(7–12).4}$ | −0.416/left-skewed | 0.028 | −0.066 | 0.056 | 0.049 | −0.005 | 0.001 | 0.914 |
| $\alpha_{(1–6)}$ | 2.269/normal | −0.080 | −0.035 | 0.174 | 0.155 | −0.006 | 0.012 | 0.895 |
| $\alpha_{(7–12)}$ | 2.269/left-skewed | −0.307 | −0.135 | 0.351 | 0.171 | 0.013 | 0.016 | 0.546 |
| $\beta_{(1–6).1}$ | −1.853/normal | −0.072 | 0.039 | 0.140 | 0.120 | | | |
| $\beta_{(1–6).2}$ | −0.850/normal | −0.003 | 0.003 | 0.065 | 0.064 | | | |
| $\beta_{(1–6).3}$ | 0.728/normal | 0.031 | 0.043 | 0.086 | 0.080 | | | |
| $\beta_{(1–6).4}$ | 2.237/normal | −0.005 | −0.002 | 0.150 | 0.150 | | | |
| $\beta_{(7–12).1}$ | −1.831/left-skewed | −0.183 | 0.100 | 0.236 | 0.149 | | | |
| $\beta_{(7–12).2}$ | −1.445/left-skewed | −0.103 | 0.071 | 0.147 | 0.106 | | | |
| $\beta_{(7–12).3}$ | −1.016/left-skewed | −0.044 | 0.044 | 0.089 | 0.077 | | | |
| $\beta_{(7–12).4}$ | −0.521/left-skewed | 0.004 | −0.008 | 0.064 | 0.063 | | | |

*Table E.49.* FA-lin standardized parameters of interest and corresponding standard errors in Cell lrnRS2 averaged over items with equal population value/shape. $n = 200$; $R = 1000$.

| | $\omega$ | PB($\hat{\omega}$) | RB($\hat{\omega}$) | RMSE($\hat{\omega}$) | SD($\hat{\omega}$) | RB($\hat{se}$) | RMSE($\hat{se}$) | 95%-cov. |
|---|---|---|---|---|---|---|---|---|
| $\lambda_{(1–4)}$ | 0.800/normal | −0.066 | −0.083 | 0.075 | 0.035 | 0.031 | 0.004 | 0.585 |
| $\lambda_{(5–8)}$ | 0.800/left-skewed | −0.304 | −0.380 | 0.306 | 0.038 | 0.493 | 0.019 | 0.000 |
| $\lambda_{(9–12)}$ | 0.800/right-skewed | −0.031 | −0.039 | 0.050 | 0.039 | −0.170 | 0.008 | 0.843 |

*Table E.50.* FA-poly standardized parameters of interest and corresponding standard errors in Cell lrnRS2 averaged over items with equal population value/shape. $n = 200$; $R = 1000$.

| | $\omega$ | PB($\hat{\omega}$) | RB($\hat{\omega}$) | RMSE($\hat{\omega}$) | SD($\hat{\omega}$) | RB($\hat{se}$) | RMSE($\hat{se}$) | 95%-cov. |
|---|---|---|---|---|---|---|---|---|
| $\lambda_{(1–4)}$ | 0.800/normal | −0.014 | −0.017 | 0.037 | 0.034 | −0.116 | 0.006 | 0.920 |
| $\lambda_{(5–8)}$ | 0.800/left-skewed | −0.091 | −0.114 | 0.102 | 0.045 | −0.021 | 0.005 | 0.453 |
| $\lambda_{(9–12)}$ | 0.800/right-skewed | 0.066 | 0.083 | 0.077 | 0.039 | −0.149 | 0.008 | 0.460 |
| $\tau_{(1–4).1}$ | −1.483/normal | −0.180 | 0.121 | 0.242 | 0.162 | −0.054 | 0.021 | 0.874 |
| $\tau_{(1–4).2}$ | −0.680/normal | 0.033 | −0.049 | 0.104 | 0.098 | −0.024 | 0.004 | 0.919 |
| $\tau_{(1–4).3}$ | 0.583/normal | 0.059 | 0.101 | 0.114 | 0.098 | −0.020 | 0.003 | 0.910 |
| $\tau_{(1–4).4}$ | 1.790/normal | −0.129 | −0.072 | 0.205 | 0.159 | −0.038 | 0.019 | 0.786 |
| $\tau_{(5–8).1}$ | −1.465/left-skewed | −0.176 | 0.120 | 0.234 | 0.154 | −0.021 | 0.017 | 0.833 |
| $\tau_{(5–8).2}$ | −1.156/left-skewed | −0.076 | 0.066 | 0.139 | 0.117 | 0.013 | 0.007 | 0.940 |
| $\tau_{(5–8).3}$ | −0.813/left-skewed | 0.009 | −0.011 | 0.100 | 0.099 | 0.009 | 0.004 | 0.954 |
| $\tau_{(5–8).4}$ | −0.416/left-skewed | 0.068 | −0.164 | 0.113 | 0.090 | 0.014 | 0.002 | 0.870 |
| $\tau_{(9–12).1}$ | 0.260/right-skewed | 0.084 | 0.324 | 0.124 | 0.091 | −0.000 | 0.002 | 0.861 |
| $\tau_{(9–12).2}$ | 0.760/right-skewed | 0.037 | 0.049 | 0.107 | 0.101 | −0.005 | 0.004 | 0.936 |
| $\tau_{(9–12).3}$ | 1.260/right-skewed | −0.037 | −0.029 | 0.128 | 0.122 | −0.033 | 0.009 | 0.922 |
| $\tau_{(9–12).4}$ | 1.760/right-skewed | −0.123 | −0.070 | 0.198 | 0.156 | −0.033 | 0.018 | 0.827 |
| $\alpha_{(1–4)}$ | 2.269/normal | −0.075 | −0.033 | 0.264 | 0.253 | −0.119 | 0.040 | 0.881 |
| $\alpha_{(5–8)}$ | 2.269/left-skewed | −0.538 | −0.237 | 0.583 | 0.223 | −0.028 | 0.029 | 0.309 |
| $\alpha_{(9–12)}$ | 2.269/right-skewed | 0.808 | 0.356 | 1.032 | 0.641 | −0.203 | 0.291 | 0.750 |
| $\beta_{(1–4).1}$ | −1.853/normal | −0.265 | 0.143 | 0.355 | 0.235 | | | |
| $\beta_{(1–4).2}$ | −0.850/normal | 0.026 | −0.030 | 0.134 | 0.132 | | | |
| $\beta_{(1–4).3}$ | 0.728/normal | 0.090 | 0.124 | 0.166 | 0.140 | | | |
| $\beta_{(1–4).4}$ | 2.237/normal | −0.118 | −0.053 | 0.282 | 0.256 | | | |
| $\beta_{(5–8).1}$ | −1.831/left-skewed | −0.497 | 0.271 | 0.574 | 0.288 | | | |
| $\beta_{(5–8).2}$ | −1.445/left-skewed | −0.302 | 0.209 | 0.372 | 0.218 | | | |
| $\beta_{(5–8).3}$ | −1.016/left-skewed | −0.124 | 0.122 | 0.209 | 0.169 | | | |
| $\beta_{(5–8).4}$ | −0.521/left-skewed | 0.027 | −0.052 | 0.135 | 0.132 | | | |
| $\beta_{(9–12).1}$ | 0.325/right-skewed | 0.073 | 0.224 | 0.128 | 0.105 | | | |
| $\beta_{(9–12).2}$ | 0.950/right-skewed | −0.027 | −0.028 | 0.133 | 0.131 | | | |
| $\beta_{(9–12).3}$ | 1.575/right-skewed | −0.159 | −0.101 | 0.235 | 0.173 | | | |
| $\beta_{(9–12).4}$ | 2.200/right-skewed | −0.304 | −0.138 | 0.378 | 0.225 | | | |

*Table E.51.* IRT-grm standardized parameters of interest and corresponding standard errors in Cell lrnRS2 averaged over items with equal population value/shape. $n = 200$; $R = 1000$.

| | $\omega$ | PB($\hat{\omega}$) | RB($\hat{\omega}$) | RMSE($\hat{\omega}$) | SD($\hat{\omega}$) | RB($\hat{se}$) | RMSE($\hat{se}$) | 95%-cov. |
|---|---|---|---|---|---|---|---|---|
| $\lambda_{(1–4)}$ | 0.800/normal | −0.012 | −0.015 | 0.039 | 0.037 | −0.018 | 0.005 | 0.950 |
| $\lambda_{(5–8)}$ | 0.800/left-skewed | −0.051 | −0.064 | 0.072 | 0.051 | −0.013 | 0.007 | 0.870 |
| $\lambda_{(9–12)}$ | 0.800/right-skewed | 0.009 | 0.011 | 0.038 | 0.037 | −0.013 | 0.006 | 0.915 |
| $\tau_{(1–4).1}$ | −1.483/normal | −0.058 | 0.039 | 0.158 | 0.147 | −0.055 | 0.023 | 0.954 |
| $\tau_{(1–4).2}$ | −0.680/normal | 0.011 | −0.016 | 0.087 | 0.087 | −0.009 | 0.004 | 0.946 |
| $\tau_{(1–4).3}$ | 0.583/normal | 0.012 | 0.020 | 0.100 | 0.100 | 0.003 | 0.004 | 0.952 |
| $\tau_{(1–4).4}$ | 1.790/normal | −0.027 | −0.015 | 0.178 | 0.176 | −0.035 | 0.022 | 0.939 |
| $\tau_{(5–8).1}$ | −1.465/left-skewed | −0.073 | 0.050 | 0.163 | 0.145 | −0.022 | 0.022 | 0.960 |
| $\tau_{(5–8).2}$ | −1.156/left-skewed | −0.018 | 0.016 | 0.106 | 0.104 | 0.017 | 0.009 | 0.963 |
| $\tau_{(5–8).3}$ | −0.813/left-skewed | 0.015 | −0.018 | 0.088 | 0.087 | 0.027 | 0.004 | 0.951 |
| $\tau_{(5–8).4}$ | −0.416/left-skewed | 0.032 | −0.078 | 0.088 | 0.082 | 0.033 | 0.003 | 0.943 |
| $\tau_{(9–12).1}$ | 0.260/right-skewed | 0.020 | 0.076 | 0.092 | 0.089 | 0.026 | 0.004 | 0.951 |
| $\tau_{(9–12).2}$ | 0.760/right-skewed | 0.001 | 0.002 | 0.103 | 0.103 | 0.026 | 0.006 | 0.954 |
| $\tau_{(9–12).3}$ | 1.260/right-skewed | −0.016 | −0.012 | 0.132 | 0.131 | −0.013 | 0.011 | 0.939 |
| $\tau_{(9–12).4}$ | 1.760/right-skewed | −0.021 | −0.012 | 0.174 | 0.173 | −0.031 | 0.021 | 0.943 |
| $\alpha_{(1–4)}$ | 2.269/normal | −0.061 | −0.027 | 0.281 | 0.275 | −0.023 | 0.035 | 0.926 |
| $\alpha_{(5–8)}$ | 2.269/left-skewed | −0.308 | −0.136 | 0.436 | 0.308 | −0.018 | 0.050 | 0.758 |
| $\alpha_{(9–12)}$ | 2.269/right-skewed | 0.116 | 0.051 | 0.347 | 0.327 | −0.027 | 0.051 | 0.952 |
| $\beta_{(1–4).1}$ | −1.853/normal | −0.108 | 0.058 | 0.250 | 0.225 | | | |
| $\beta_{(1–4).2}$ | −0.850/normal | −0.001 | 0.001 | 0.115 | 0.115 | | | |
| $\beta_{(1–4).3}$ | 0.728/normal | 0.029 | 0.040 | 0.144 | 0.141 | | | |
| $\beta_{(1–4).4}$ | 2.237/normal | 0.008 | 0.004 | 0.276 | 0.276 | | | |
| $\beta_{(5–8).1}$ | −1.831/left-skewed | −0.237 | 0.130 | 0.362 | 0.273 | | | |
| $\beta_{(5–8).2}$ | −1.445/left-skewed | −0.133 | 0.092 | 0.235 | 0.194 | | | |
| $\beta_{(5–8).3}$ | −1.016/left-skewed | −0.055 | 0.055 | 0.150 | 0.140 | | | |
| $\beta_{(5–8).4}$ | −0.521/left-skewed | 0.007 | −0.013 | 0.110 | 0.110 | | | |
| $\beta_{(9–12).1}$ | 0.325/right-skewed | 0.022 | 0.067 | 0.116 | 0.114 | | | |
| $\beta_{(9–12).2}$ | 0.950/right-skewed | −0.006 | −0.006 | 0.144 | 0.144 | | | |
| $\beta_{(9–12).3}$ | 1.575/right-skewed | −0.032 | −0.020 | 0.198 | 0.196 | | | |
| $\beta_{(9–12).4}$ | 2.200/right-skewed | −0.044 | −0.020 | 0.265 | 0.262 | | | |

*Table E.52.* FA-lin standardized parameters of interest and corresponding standard errors in Cell lrnRS6 averaged over items with equal population value/shape. $n = 600$; $R = 1000$.

| | $\omega$ | PB($\hat{\omega}$) | RB($\hat{\omega}$) | RMSE($\hat{\omega}$) | SD($\hat{\omega}$) | RB($\hat{se}$) | RMSE($\hat{se}$) | 95%-cov. |
|---|---|---|---|---|---|---|---|---|
| $\lambda_{(1–4)}$ | 0.800/normal | −0.067 | −0.083 | 0.070 | 0.020 | 0.049 | 0.002 | 0.062 |
| $\lambda_{(5–8)}$ | 0.800/left-skewed | −0.305 | −0.382 | 0.306 | 0.022 | 0.474 | 0.011 | 0.000 |
| $\lambda_{(9–12)}$ | 0.800/right-skewed | −0.030 | −0.038 | 0.038 | 0.023 | −0.187 | 0.005 | 0.638 |

*Table E.53.* FA-poly standardized parameters of interest and corresponding standard errors in Cell lrnRS6 averaged over items with equal population value/shape. $n = 600$; $R = 1000$.

| | $\omega$ | PB($\hat{\omega}$) | RB($\hat{\omega}$) | RMSE($\hat{\omega}$) | SD($\hat{\omega}$) | RB($\hat{se}$) | RMSE($\hat{se}$) | 95%-cov. |
|---|---|---|---|---|---|---|---|---|
| $\lambda_{(1–4)}$ | 0.800/normal | −0.016 | −0.020 | 0.025 | 0.020 | −0.081 | 0.002 | 0.863 |
| $\lambda_{(5–8)}$ | 0.800/left-skewed | −0.097 | −0.121 | 0.100 | 0.026 | 0.026 | 0.002 | 0.019 |
| $\lambda_{(9–12)}$ | 0.800/right-skewed | 0.064 | 0.080 | 0.068 | 0.022 | −0.116 | 0.003 | 0.164 |
| $\tau_{(1–4).1}$ | −1.483/normal | −0.172 | 0.116 | 0.193 | 0.088 | −0.007 | 0.005 | 0.516 |
| $\tau_{(1–4).2}$ | −0.680/normal | 0.034 | −0.050 | 0.066 | 0.056 | −0.014 | 0.001 | 0.893 |
| $\tau_{(1–4).3}$ | 0.583/normal | 0.061 | 0.105 | 0.083 | 0.057 | −0.031 | 0.002 | 0.802 |
| $\tau_{(1–4).4}$ | 1.790/normal | −0.139 | −0.078 | 0.165 | 0.089 | −0.019 | 0.006 | 0.620 |
| $\tau_{(5–8).1}$ | −1.465/left-skewed | −0.162 | 0.111 | 0.184 | 0.086 | −0.008 | 0.005 | 0.561 |
| $\tau_{(5–8).2}$ | −1.156/left-skewed | −0.068 | 0.059 | 0.096 | 0.069 | −0.008 | 0.003 | 0.855 |
| $\tau_{(5–8).3}$ | −0.813/left-skewed | 0.012 | −0.014 | 0.060 | 0.059 | −0.025 | 0.002 | 0.935 |
| $\tau_{(5–8).4}$ | −0.416/left-skewed | 0.069 | −0.166 | 0.087 | 0.053 | −0.017 | 0.001 | 0.714 |
| $\tau_{(9–12).1}$ | 0.260/right-skewed | 0.086 | 0.329 | 0.101 | 0.053 | −0.008 | 0.001 | 0.616 |
| $\tau_{(9–12).2}$ | 0.760/right-skewed | 0.040 | 0.053 | 0.072 | 0.060 | −0.035 | 0.002 | 0.889 |
| $\tau_{(9–12).3}$ | 1.260/right-skewed | −0.038 | −0.030 | 0.080 | 0.071 | −0.043 | 0.004 | 0.900 |
| $\tau_{(9–12).4}$ | 1.760/right-skewed | −0.134 | −0.076 | 0.161 | 0.090 | −0.043 | 0.007 | 0.642 |
| $\alpha_{(1–4)}$ | 2.269/normal | −0.111 | −0.049 | 0.179 | 0.141 | −0.080 | 0.014 | 0.814 |
| $\alpha_{(5–8)}$ | 2.269/left-skewed | −0.578 | −0.255 | 0.591 | 0.122 | 0.022 | 0.009 | 0.012 |
| $\alpha_{(9–12)}$ | 2.269/right-skewed | 0.692 | 0.305 | 0.759 | 0.313 | −0.125 | 0.062 | 0.248 |
| $\beta_{(1–4).1}$ | −1.853/normal | −0.258 | 0.139 | 0.288 | 0.129 | | | |
| $\beta_{(1–4).2}$ | −0.850/normal | 0.026 | −0.030 | 0.079 | 0.075 | | | |
| $\beta_{(1–4).3}$ | 0.728/normal | 0.094 | 0.128 | 0.124 | 0.081 | | | |
| $\beta_{(1–4).4}$ | 2.237/normal | −0.129 | −0.058 | 0.192 | 0.142 | | | |
| $\beta_{(5–8).1}$ | −1.831/left-skewed | −0.486 | 0.265 | 0.512 | 0.162 | | | |
| $\beta_{(5–8).2}$ | −1.445/left-skewed | −0.298 | 0.206 | 0.323 | 0.125 | | | |
| $\beta_{(5–8).3}$ | −1.016/left-skewed | −0.125 | 0.123 | 0.159 | 0.099 | | | |
| $\beta_{(5–8).4}$ | −0.521/left-skewed | 0.026 | −0.050 | 0.083 | 0.078 | | | |
| $\beta_{(9–12).1}$ | 0.325/right-skewed | 0.075 | 0.231 | 0.097 | 0.061 | | | |
| $\beta_{(9–12).2}$ | 0.950/right-skewed | −0.023 | −0.024 | 0.081 | 0.078 | | | |
| $\beta_{(9–12).3}$ | 1.575/right-skewed | −0.159 | −0.101 | 0.188 | 0.101 | | | |
| $\beta_{(9–12).4}$ | 2.200/right-skewed | −0.316 | −0.144 | 0.342 | 0.130 | | | |

*Table E.54.* IRT-grm standardized parameters of interest and corresponding standard errors in Cell lrnRS6 averaged over items with equal population value/shape. $n = 600$; $R = 1000$.

| | $\omega$ | PB($\hat{\omega}$) | RB($\hat{\omega}$) | RMSE($\hat{\omega}$) | SD($\hat{\omega}$) | RB($\hat{se}$) | RMSE($\hat{se}$) | 95%-cov. |
|---|---|---|---|---|---|---|---|---|
| $\lambda_{(1–4)}$ | 0.800/normal | −0.013 | −0.017 | 0.025 | 0.021 | 0.002 | 0.002 | 0.922 |
| $\lambda_{(5–8)}$ | 0.800/left-skewed | −0.052 | −0.065 | 0.059 | 0.029 | −0.003 | 0.002 | 0.591 |
| $\lambda_{(9–12)}$ | 0.800/right-skewed | 0.008 | 0.010 | 0.023 | 0.022 | −0.012 | 0.002 | 0.918 |
| $\tau_{(1–4).1}$ | −1.483/normal | −0.050 | 0.033 | 0.094 | 0.079 | −0.011 | 0.006 | 0.925 |
| $\tau_{(1–4).2}$ | −0.680/normal | 0.013 | −0.019 | 0.052 | 0.051 | −0.022 | 0.001 | 0.938 |
| $\tau_{(1–4).3}$ | 0.583/normal | 0.013 | 0.022 | 0.060 | 0.059 | −0.018 | 0.002 | 0.944 |
| $\tau_{(1–4).4}$ | 1.790/normal | −0.039 | −0.022 | 0.105 | 0.098 | −0.011 | 0.006 | 0.917 |
| $\tau_{(5–8).1}$ | −1.465/left-skewed | −0.059 | 0.040 | 0.100 | 0.081 | −0.013 | 0.007 | 0.909 |
| $\tau_{(5–8).2}$ | −1.156/left-skewed | −0.010 | 0.008 | 0.062 | 0.061 | −0.010 | 0.003 | 0.947 |
| $\tau_{(5–8).3}$ | −0.813/left-skewed | 0.018 | −0.023 | 0.056 | 0.053 | −0.028 | 0.002 | 0.928 |
| $\tau_{(5–8).4}$ | −0.416/left-skewed | 0.034 | −0.083 | 0.061 | 0.050 | −0.022 | 0.001 | 0.883 |
| $\tau_{(9–12).1}$ | 0.260/right-skewed | 0.020 | 0.078 | 0.057 | 0.053 | −0.000 | 0.002 | 0.931 |
| $\tau_{(9–12).2}$ | 0.760/right-skewed | 0.003 | 0.004 | 0.062 | 0.062 | −0.014 | 0.002 | 0.948 |
| $\tau_{(9–12).3}$ | 1.260/right-skewed | −0.017 | −0.014 | 0.079 | 0.077 | −0.028 | 0.004 | 0.929 |
| $\tau_{(9–12).4}$ | 1.760/right-skewed | −0.034 | −0.019 | 0.105 | 0.099 | −0.034 | 0.008 | 0.919 |
| $\alpha_{(1–4)}$ | 2.269/normal | −0.090 | −0.040 | 0.176 | 0.151 | 0.003 | 0.011 | 0.890 |
| $\alpha_{(5–8)}$ | 2.269/left-skewed | −0.337 | −0.149 | 0.378 | 0.172 | −0.008 | 0.016 | 0.473 |
| $\alpha_{(9–12)}$ | 2.269/right-skewed | 0.076 | 0.034 | 0.197 | 0.181 | −0.013 | 0.015 | 0.947 |
| $\beta_{(1–4).1}$ | −1.853/normal | −0.096 | 0.052 | 0.156 | 0.123 | | | |
| $\beta_{(1–4).2}$ | −0.850/normal | 0.002 | −0.002 | 0.066 | 0.066 | | | |
| $\beta_{(1–4).3}$ | 0.728/normal | 0.030 | 0.041 | 0.087 | 0.082 | | | |
| $\beta_{(1–4).4}$ | 2.237/normal | −0.009 | −0.004 | 0.152 | 0.152 | | | |
| $\beta_{(5–8).1}$ | −1.831/left-skewed | −0.209 | 0.114 | 0.258 | 0.151 | | | |
| $\beta_{(5–8).2}$ | −1.445/left-skewed | −0.116 | 0.080 | 0.158 | 0.108 | | | |
| $\beta_{(5–8).3}$ | −1.016/left-skewed | −0.047 | 0.047 | 0.094 | 0.081 | | | |
| $\beta_{(5–8).4}$ | −0.521/left-skewed | 0.010 | −0.019 | 0.067 | 0.066 | | | |
| $\beta_{(9–12).1}$ | 0.325/right-skewed | 0.022 | 0.069 | 0.071 | 0.067 | | | |
| $\beta_{(9–12).2}$ | 0.950/right-skewed | −0.004 | −0.005 | 0.086 | 0.086 | | | |
| $\beta_{(9–12).3}$ | 1.575/right-skewed | −0.035 | −0.022 | 0.119 | 0.114 | | | |
| $\beta_{(9–12).4}$ | 2.200/right-skewed | −0.061 | −0.028 | 0.163 | 0.151 | | | |

*Table E.55.* FA-lin standardized parameters of interest and corresponding standard errors in Cell bnRS2 averaged over items with equal population value/shape. $n = 200$; $R = 1000$.

| | $\omega$ | PB($\hat{\omega}$) | RB($\hat{\omega}$) | RMSE($\hat{\omega}$) | SD($\hat{\omega}$) | RB($\hat{se}$) | RMSE($\hat{se}$) | 95%-cov. |
|---|---|---|---|---|---|---|---|---|
| $\lambda_{(1–6)}$ | 0.800/normal | −0.068 | −0.085 | 0.076 | 0.035 | 0.015 | 0.004 | 0.546 |
| $\lambda_{(7–12)}$ | 0.800/bimodal | −0.065 | −0.081 | 0.074 | 0.035 | 0.004 | 0.004 | 0.575 |

*Table E.56.* FA-poly standardized parameters of interest and corresponding standard errors in Cell bnRS2 averaged over items with equal population value/shape. $n = 200$; $R = 1000$.

| | $\omega$ | PB($\hat{\omega}$) | RB($\hat{\omega}$) | RMSE($\hat{\omega}$) | SD($\hat{\omega}$) | RB($\hat{se}$) | RMSE($\hat{se}$) | 95%-cov. |
|---|---|---|---|---|---|---|---|---|
| $\lambda_{(1–6)}$ | 0.800/normal | −0.019 | −0.024 | 0.039 | 0.034 | −0.063 | 0.004 | 0.922 |
| $\lambda_{(7–12)}$ | 0.800/bimodal | −0.017 | −0.021 | 0.037 | 0.034 | −0.055 | 0.004 | 0.926 |
| $\tau_{(1–6).1}$ | −1.483/normal | −0.183 | 0.123 | 0.242 | 0.159 | −0.035 | 0.019 | 0.872 |
| $\tau_{(1–6).2}$ | −0.680/normal | 0.034 | −0.049 | 0.103 | 0.097 | −0.010 | 0.003 | 0.924 |
| $\tau_{(1–6).3}$ | 0.583/normal | 0.065 | 0.111 | 0.116 | 0.096 | −0.002 | 0.003 | 0.913 |
| $\tau_{(1–6).4}$ | 1.790/normal | −0.124 | −0.069 | 0.200 | 0.157 | −0.021 | 0.019 | 0.802 |
| $\tau_{(7–12).1}$ | −1.203/bimodal | −0.087 | 0.072 | 0.154 | 0.127 | −0.039 | 0.010 | 0.902 |
| $\tau_{(7–12).2}$ | −0.212/bimodal | 0.086 | −0.405 | 0.123 | 0.089 | 0.003 | 0.001 | 0.822 |
| $\tau_{(7–12).3}$ | 0.034/bimodal | 0.093 | 2.735 | 0.128 | 0.089 | 0.003 | 0.001 | 0.821 |
| $\tau_{(7–12).4}$ | 1.334/bimodal | −0.040 | −0.030 | 0.130 | 0.124 | −0.015 | 0.008 | 0.923 |
| $\alpha_{(1–6)}$ | 2.269/normal | −0.115 | −0.051 | 0.267 | 0.241 | −0.062 | 0.031 | 0.875 |
| $\alpha_{(7–12)}$ | 2.269/bimodal | −0.101 | −0.045 | 0.262 | 0.242 | −0.057 | 0.030 | 0.885 |
| $\beta_{(1–6).1}$ | −1.853/normal | −0.283 | 0.153 | 0.366 | 0.231 | | | |
| $\beta_{(1–6).2}$ | −0.850/normal | 0.021 | −0.025 | 0.132 | 0.131 | | | |
| $\beta_{(1–6).3}$ | 0.728/normal | 0.103 | 0.142 | 0.173 | 0.139 | | | |
| $\beta_{(1–6).4}$ | 2.237/normal | −0.097 | −0.043 | 0.271 | 0.253 | | | |
| $\beta_{(7–12).1}$ | −1.504/bimodal | −0.147 | 0.098 | 0.237 | 0.187 | | | |
| $\beta_{(7–12).2}$ | −0.265/bimodal | 0.104 | −0.391 | 0.154 | 0.114 | | | |
| $\beta_{(7–12).3}$ | 0.042/bimodal | 0.120 | 2.837 | 0.166 | 0.115 | | | |
| $\beta_{(7–12).4}$ | 1.667/bimodal | −0.010 | −0.006 | 0.198 | 0.198 | | | |

*Table E.57.* IRT-grm standardized parameters of interest and corresponding standard errors in Cell bnRS2 averaged over items with equal population value/shape. $n = 200$; $R = 1000$.

| | $\omega$ | PB($\hat{\omega}$) | RB($\hat{\omega}$) | RMSE($\hat{\omega}$) | SD($\hat{\omega}$) | RB($\hat{se}$) | RMSE($\hat{se}$) | 95%-cov. |
|---|---|---|---|---|---|---|---|---|
| $\lambda_{(1–6)}$ | 0.800/normal | −0.012 | −0.014 | 0.038 | 0.036 | −0.008 | 0.005 | 0.949 |
| $\lambda_{(7–12)}$ | 0.800/bimodal | −0.016 | −0.020 | 0.039 | 0.036 | 0.006 | 0.004 | 0.948 |
| $\tau_{(1–6).1}$ | −1.483/normal | −0.039 | 0.026 | 0.146 | 0.141 | −0.031 | 0.021 | 0.960 |
| $\tau_{(1–6).2}$ | −0.680/normal | 0.011 | −0.016 | 0.084 | 0.083 | 0.009 | 0.003 | 0.948 |
| $\tau_{(1–6).3}$ | 0.583/normal | 0.009 | 0.016 | 0.101 | 0.100 | 0.012 | 0.005 | 0.953 |
| $\tau_{(1–6).4}$ | 1.790/normal | 0.001 | 0.001 | 0.178 | 0.178 | −0.018 | 0.022 | 0.952 |
| $\tau_{(7–12).1}$ | −1.203/bimodal | −0.007 | 0.006 | 0.109 | 0.109 | −0.032 | 0.010 | 0.947 |
| $\tau_{(7–12).2}$ | −0.212/bimodal | 0.016 | −0.076 | 0.083 | 0.081 | 0.021 | 0.002 | 0.950 |
| $\tau_{(7–12).3}$ | 0.034/bimodal | 0.014 | 0.403 | 0.086 | 0.085 | 0.021 | 0.003 | 0.954 |
| $\tau_{(7–12).4}$ | 1.334/bimodal | 0.007 | 0.005 | 0.139 | 0.139 | −0.003 | 0.011 | 0.954 |
| $\alpha_{(1–6)}$ | 2.269/normal | −0.058 | −0.025 | 0.275 | 0.269 | −0.009 | 0.034 | 0.928 |
| $\alpha_{(7–12)}$ | 2.269/bimodal | −0.092 | −0.041 | 0.277 | 0.261 | 0.001 | 0.033 | 0.911 |
| $\beta_{(1–6).1}$ | −1.853/normal | −0.082 | 0.044 | 0.233 | 0.218 | | | |
| $\beta_{(1–6).2}$ | −0.850/normal | 0.000 | −0.000 | 0.110 | 0.110 | | | |
| $\beta_{(1–6).3}$ | 0.728/normal | 0.026 | 0.035 | 0.144 | 0.142 | | | |
| $\beta_{(1–6).4}$ | 2.237/normal | 0.042 | 0.019 | 0.278 | 0.275 | | | |
| $\beta_{(7–12).1}$ | −1.504/bimodal | −0.044 | 0.029 | 0.174 | 0.168 | | | |
| $\beta_{(7–12).2}$ | −0.265/bimodal | 0.015 | −0.058 | 0.104 | 0.103 | | | |
| $\beta_{(7–12).3}$ | 0.042/bimodal | 0.020 | 0.465 | 0.112 | 0.110 | | | |
| $\beta_{(7–12).4}$ | 1.667/bimodal | 0.048 | 0.029 | 0.220 | 0.215 | | | |

*Table E.58.* FA-lin standardized parameters of interest and corresponding standard errors in Cell bnRS6 averaged over items with equal population value/shape. $n = 600$; $R = 1000$.

| | $\omega$ | PB($\hat{\omega}$) | RB($\hat{\omega}$) | RMSE($\hat{\omega}$) | SD($\hat{\omega}$) | RB($\hat{se}$) | RMSE($\hat{se}$) | 95%-cov. |
|---|---|---|---|---|---|---|---|---|
| $\lambda_{(1–6)}$ | 0.800/normal | −0.066 | −0.082 | 0.069 | 0.020 | 0.034 | 0.001 | 0.060 |
| $\lambda_{(7–12)}$ | 0.800/bimodal | −0.064 | −0.080 | 0.067 | 0.020 | 0.000 | 0.001 | 0.079 |

*Table E.59.* FA-poly standardized parameters of interest and corresponding standard errors in Cell bnRS6 averaged over items with equal population value/shape. $n = 600$; $R = 1000$.

| | $\omega$ | PB($\hat{\omega}$) | RB($\hat{\omega}$) | RMSE($\hat{\omega}$) | SD($\hat{\omega}$) | RB($\hat{se}$) | RMSE($\hat{se}$) | 95%-cov. |
|---|---|---|---|---|---|---|---|---|
| $\lambda_{(1–6)}$ | 0.800/normal | −0.019 | −0.024 | 0.027 | 0.019 | −0.014 | 0.001 | 0.851 |
| $\lambda_{(7–12)}$ | 0.800/bimodal | −0.018 | −0.022 | 0.026 | 0.019 | −0.033 | 0.001 | 0.861 |
| $\tau_{(1–6).1}$ | −1.483/normal | −0.169 | 0.114 | 0.190 | 0.086 | 0.009 | 0.006 | 0.521 |
| $\tau_{(1–6).2}$ | −0.680/normal | 0.036 | −0.053 | 0.066 | 0.055 | 0.007 | 0.001 | 0.894 |
| $\tau_{(1–6).3}$ | 0.583/normal | 0.061 | 0.105 | 0.082 | 0.054 | 0.014 | 0.001 | 0.809 |
| $\tau_{(1–6).4}$ | 1.790/normal | −0.137 | −0.076 | 0.162 | 0.087 | 0.001 | 0.005 | 0.629 |
| $\tau_{(7–12).1}$ | −1.203/bimodal | −0.082 | 0.068 | 0.109 | 0.071 | −0.015 | 0.003 | 0.804 |
| $\tau_{(7–12).2}$ | −0.212/bimodal | 0.085 | −0.403 | 0.100 | 0.051 | 0.005 | 0.001 | 0.617 |
| $\tau_{(7–12).3}$ | 0.034/bimodal | 0.092 | 2.709 | 0.105 | 0.052 | −0.007 | 0.001 | 0.581 |
| $\tau_{(7–12).4}$ | 1.334/bimodal | −0.049 | −0.037 | 0.085 | 0.070 | 0.004 | 0.003 | 0.879 |
| $\alpha_{(1–6)}$ | 2.269/normal | −0.133 | −0.059 | 0.189 | 0.134 | −0.016 | 0.009 | 0.797 |
| $\alpha_{(7–12)}$ | 2.269/bimodal | −0.125 | −0.055 | 0.186 | 0.137 | −0.031 | 0.010 | 0.811 |
| $\beta_{(1–6).1}$ | −1.853/normal | −0.263 | 0.142 | 0.292 | 0.126 | | | |
| $\beta_{(1–6).2}$ | −0.850/normal | 0.025 | −0.029 | 0.078 | 0.074 | | | |
| $\beta_{(1–6).3}$ | 0.728/normal | 0.097 | 0.134 | 0.125 | 0.078 | | | |
| $\beta_{(1–6).4}$ | 2.237/normal | −0.118 | −0.053 | 0.183 | 0.139 | | | |
| $\beta_{(7–12).1}$ | −1.504/bimodal | −0.141 | 0.094 | 0.175 | 0.104 | | | |
| $\beta_{(7–12).2}$ | −0.265/bimodal | 0.103 | −0.390 | 0.122 | 0.065 | | | |
| $\beta_{(7–12).3}$ | 0.042/bimodal | 0.118 | 2.800 | 0.136 | 0.067 | | | |
| $\beta_{(7–12).4}$ | 1.667/bimodal | −0.023 | −0.014 | 0.113 | 0.111 | | | |

*Table E.60.* IRT-grm standardized parameters of interest and corresponding standard errors in Cell bnRS6 averaged over items with equal population value/shape. $n = 600$; $R = 1000$.

| | $\omega$ | PB($\hat{\omega}$) | RB($\hat{\omega}$) | RMSE($\hat{\omega}$) | SD($\hat{\omega}$) | RB($\hat{se}$) | RMSE($\hat{se}$) | 95%-cov. |
|---|---|---|---|---|---|---|---|---|
| $\lambda_{(1–6)}$ | 0.800/normal | −0.011 | −0.013 | 0.023 | 0.020 | 0.006 | 0.002 | 0.939 |
| $\lambda_{(7–12)}$ | 0.800/bimodal | −0.016 | −0.020 | 0.026 | 0.021 | 0.002 | 0.002 | 0.907 |
| $\tau_{(1–6).1}$ | −1.483/normal | −0.025 | 0.017 | 0.081 | 0.077 | 0.006 | 0.006 | 0.948 |
| $\tau_{(1–6).2}$ | −0.680/normal | 0.013 | −0.019 | 0.050 | 0.048 | 0.005 | 0.001 | 0.945 |
| $\tau_{(1–6).3}$ | 0.583/normal | 0.007 | 0.012 | 0.058 | 0.057 | 0.019 | 0.002 | 0.956 |
| $\tau_{(1–6).4}$ | 1.790/normal | −0.014 | −0.008 | 0.100 | 0.099 | 0.007 | 0.006 | 0.950 |
| $\tau_{(7–12).1}$ | −1.203/bimodal | −0.003 | 0.002 | 0.061 | 0.061 | −0.014 | 0.003 | 0.949 |
| $\tau_{(7–12).2}$ | −0.212/bimodal | 0.015 | −0.073 | 0.050 | 0.048 | 0.005 | 0.001 | 0.934 |
| $\tau_{(7–12).3}$ | 0.034/bimodal | 0.013 | 0.374 | 0.052 | 0.050 | −0.005 | 0.001 | 0.941 |
| $\tau_{(7–12).4}$ | 1.334/bimodal | −0.003 | −0.002 | 0.078 | 0.078 | 0.011 | 0.004 | 0.951 |
| $\alpha_{(1–6)}$ | 2.269/normal | −0.071 | −0.031 | 0.167 | 0.151 | 0.004 | 0.011 | 0.910 |
| $\alpha_{(7–12)}$ | 2.269/bimodal | −0.110 | −0.048 | 0.184 | 0.148 | 0.003 | 0.011 | 0.864 |
| $\beta_{(1–6).1}$ | −1.853/normal | −0.059 | 0.032 | 0.133 | 0.119 | | | |
| $\beta_{(1–6).2}$ | −0.850/normal | 0.004 | −0.005 | 0.063 | 0.063 | | | |
| $\beta_{(1–6).3}$ | 0.728/normal | 0.019 | 0.027 | 0.083 | 0.080 | | | |
| $\beta_{(1–6).4}$ | 2.237/normal | 0.015 | 0.007 | 0.151 | 0.151 | | | |
| $\beta_{(7–12).1}$ | −1.504/bimodal | −0.035 | 0.023 | 0.100 | 0.094 | | | |
| $\beta_{(7–12).2}$ | −0.265/bimodal | 0.014 | −0.054 | 0.062 | 0.060 | | | |
| $\beta_{(7–12).3}$ | 0.042/bimodal | 0.017 | 0.412 | 0.067 | 0.065 | | | |
| $\beta_{(7–12).4}$ | 1.667/bimodal | 0.032 | 0.019 | 0.124 | 0.120 | | | |

### E.1.2 Coverage Results for $\lambda$ and $\tau$

Coverage rates of loading $\lambda$ and threshold $\tau$ parameter estimators are presented for each cell of the design. In Tables E.61 and E.62 the results are shown for the normal LV conditions for the small and medium sample size, respectively. Analogously, Tables E.63 and E.64 contain the results for the skew-normal LV conditions. These coverage rates are averaged over parameters of equally shaped items with equal population values. As a result, the coverage rates are based on various numbers of replications. For example, the average coverage rate of $\lambda_{1–12}$ in Cell nNS2 is based on $R = 12000$ replications, as all these items are normally distributed; the average coverage rate of $\lambda_{1–6}$ in Cell rnNS2, however, is based on $R = 6000$ replications. The values presented in the tables are deviations from the expected coverage rate, i.e., 0.95 is subtracted from the empirical average. Coverage rates are considered acceptable when they are between 0.90 and 0.98, corresponding to a deviation between −0.05 and 0.03. Coverage rates beyond these limits are printed in boldface.

*Table E.61.* Coverage rate of 95%-confidence interval estimators for Cells nNS2, rnNS2, lnNS2, lrnNS2, and bnNS2 averaged over parameters with equal population value/shape. $R = 1000$ per parameter. Numbers represent deviation from 0.95, with values outside the range $(-0.05, 0.03)$ printed in boldface, indicating an unacceptable coverage rate.

| Cell | | $\omega$ | FA-lin | FA-poly | IRT-grm |
|---|---|---|---|---|---|
| nNS2 | $\lambda_{(1–12)}$ | 0.800/normal | **−0.170** | −0.023 | −0.004 |
| | $\tau_{(1–12).1}$ | −1.645/normal | | −0.001 | 0.002 |
| | $\tau_{(1–12).2}$ | −0.643/normal | | −0.002 | −0.002 |
| | $\tau_{(1–12).3}$ | 0.643/normal | | −0.003 | −0.005 |
| | $\tau_{(1–12).4}$ | 1.645/normal | | −0.003 | −0.002 |
| rnNS2 | $\lambda_{(1–6)}$ | 0.800/normal | **−0.312** | −0.032 | −0.008 |
| | $\lambda_{(7–12)}$ | 0.800/right-skewed | **−0.658** | −0.032 | −0.006 |
| | $\tau_{(1–6).1}$ | −1.645/normal | | −0.001 | 0.004 |
| | $\tau_{(1–6).2}$ | −0.643/normal | | −0.004 | −0.004 |
| | $\tau_{(1–6).3}$ | 0.643/normal | | −0.006 | −0.005 |
| | $\tau_{(1–6).4}$ | 1.645/normal | | 0.004 | 0.003 |
| | $\tau_{(7–12).1}$ | 0.346/right-skewed | | −0.004 | −0.007 |
| | $\tau_{(7–12).2}$ | 0.800/right-skewed | | −0.012 | −0.012 |
| | $\tau_{(7–12).3}$ | 1.221/right-skewed | | 0.003 | −0.009 |
| | $\tau_{(7–12).4}$ | 1.622/right-skewed | | −0.007 | −0.003 |
| lnNS2 | $\lambda_{(1–6)}$ | 0.800/normal | **−0.339** | −0.025 | −0.007 |
| | $\lambda_{(7–12)}$ | 0.800/left-skewed | **−0.670** | −0.021 | 0.002 |
| | $\tau_{(1–6).1}$ | −1.645/normal | | 0.002 | 0.006 |
| | $\tau_{(1–6).2}$ | −0.643/normal | | −0.002 | 0.001 |
| | $\tau_{(1–6).3}$ | 0.643/normal | | −0.008 | −0.008 |
| | $\tau_{(1–6).4}$ | 1.645/normal | | −0.005 | 0.001 |
| | $\tau_{(7–12).1}$ | −1.622/left-skewed | | 0.001 | 0.007 |
| | $\tau_{(7–12).2}$ | −1.221/left-skewed | | 0.007 | −0.005 |
| | $\tau_{(7–12).3}$ | −0.800/left-skewed | | −0.005 | −0.002 |
| | $\tau_{(7–12).4}$ | −0.346/left-skewed | | 0.003 | 0.001 |
| lrnNS2 | $\lambda_{(1–4)}$ | 0.800/normal | **−0.125** | −0.020 | 0.000 |
| | $\lambda_{(5–8)}$ | 0.800/left-skewed | **−0.918** | −0.032 | −0.008 |
| | $\lambda_{(9–12)}$ | 0.800/right-skewed | **−0.912** | −0.033 | −0.002 |
| | $\tau_{(1–4).1}$ | −1.645/normal | | −0.000 | −0.002 |
| | $\tau_{(1–4).2}$ | −0.643/normal | | −0.010 | −0.006 |
| | $\tau_{(1–4).3}$ | 0.643/normal | | −0.006 | −0.005 |
| | $\tau_{(1–4).4}$ | 1.645/normal | | −0.001 | 0.002 |
| | $\tau_{(5–8).1}$ | −1.622/left-skewed | | −0.006 | 0.000 |
| | $\tau_{(5–8).2}$ | −1.221/left-skewed | | 0.004 | −0.007 |
| | $\tau_{(5–8).3}$ | −0.800/left-skewed | | −0.008 | −0.009 |
| | $\tau_{(5–8).4}$ | −0.346/left-skewed | | 0.000 | −0.002 |
| | $\tau_{(9–12).1}$ | 0.346/right-skewed | | 0.006 | −0.004 |
| | $\tau_{(9–12).2}$ | 0.800/right-skewed | | −0.003 | −0.007 |
| | $\tau_{(9–12).3}$ | 1.221/right-skewed | | 0.008 | −0.006 |
| | $\tau_{(9–12).4}$ | 1.622/right-skewed | | −0.004 | 0.000 |
| bnNS2 | $\lambda_{(1–6)}$ | 0.800/normal | **−0.187** | −0.028 | −0.005 |
| | $\lambda_{(7–12)}$ | 0.800/bimodal | **−0.212** | −0.028 | −0.004 |
| | $\tau_{(1–6).1}$ | −1.645/normal | | −0.001 | 0.002 |
| | $\tau_{(1–6).2}$ | −0.643/normal | | 0.002 | −0.001 |
| | $\tau_{(1–6).3}$ | 0.643/normal | | −0.004 | −0.006 |
| | $\tau_{(1–6).4}$ | 1.645/normal | | 0.001 | 0.001 |
| | $\tau_{(7–12).1}$ | −1.282/bimodal | | 0.007 | 0.000 |
| | $\tau_{(7–12).2}$ | −0.126/bimodal | | −0.006 | −0.007 |
| | $\tau_{(7–12).3}$ | 0.126/bimodal | | −0.005 | −0.003 |
| | $\tau_{(7–12).4}$ | 1.282/bimodal | | 0.007 | 0.000 |

*Table E.62.* Coverage rate of 95%-confidence interval estimators for Cells nNS6, rnNS6, lnNS6, lrnNS6, and bnNS6 averaged over parameters with equal population value/shape. $R = 1000$ per parameter. Numbers represent deviation from 0.95, with values outside the range $(-0.05, 0.03)$ printed in boldface, indicating an unacceptable coverage rate.

| Cell | | $\omega$ | FA-lin | FA-poly | IRT-grm |
|---|---|---|---|---|---|
| nNS6 | $\lambda_{(1–12)}$ | 0.800/normal | **−0.597** | −0.007 | 0.002 |
| | $\tau_{(1–12).1}$ | −1.645/normal | | 0.001 | −0.002 |
| | $\tau_{(1–12).2}$ | −0.643/normal | | −0.006 | −0.009 |
| | $\tau_{(1–12).3}$ | 0.643/normal | | −0.004 | −0.005 |
| | $\tau_{(1–12).4}$ | 1.645/normal | | 0.000 | −0.003 |
| rnNS6 | $\lambda_{(1–6)}$ | 0.800/normal | **−0.837** | −0.011 | 0.000 |
| | $\lambda_{(7–12)}$ | 0.800/right-skewed | **−0.945** | −0.014 | −0.005 |
| | $\tau_{(1–6).1}$ | −1.645/normal | | −0.003 | −0.004 |
| | $\tau_{(1–6).2}$ | −0.643/normal | | −0.004 | −0.005 |
| | $\tau_{(1–6).3}$ | 0.643/normal | | −0.002 | −0.005 |
| | $\tau_{(1–6).4}$ | 1.645/normal | | 0.001 | −0.004 |
| | $\tau_{(7–12).1}$ | 0.346/right-skewed | | 0.001 | −0.002 |
| | $\tau_{(7–12).2}$ | 0.800/right-skewed | | −0.004 | −0.011 |
| | $\tau_{(7–12).3}$ | 1.221/right-skewed | | −0.005 | −0.013 |
| | $\tau_{(7–12).4}$ | 1.622/right-skewed | | −0.009 | −0.009 |
| lnNS6 | $\lambda_{(1–6)}$ | 0.800/normal | **−0.841** | −0.010 | 0.001 |
| | $\lambda_{(7–12)}$ | 0.800/left-skewed | **−0.941** | −0.022 | −0.006 |
| | $\tau_{(1–6).1}$ | −1.645/normal | | 0.002 | −0.003 |
| | $\tau_{(1–6).2}$ | −0.643/normal | | 0.002 | −0.003 |
| | $\tau_{(1–6).3}$ | 0.643/normal | | 0.006 | −0.003 |
| | $\tau_{(1–6).4}$ | 1.645/normal | | −0.001 | −0.002 |
| | $\tau_{(7–12).1}$ | −1.622/left-skewed | | −0.004 | −0.006 |
| | $\tau_{(7–12).2}$ | −1.221/left-skewed | | 0.000 | −0.008 |
| | $\tau_{(7–12).3}$ | −0.800/left-skewed | | 0.002 | −0.012 |
| | $\tau_{(7–12).4}$ | −0.346/left-skewed | | 0.006 | 0.003 |
| lrnNS6 | $\lambda_{(1–4)}$ | 0.800/normal | **−0.498** | −0.007 | −0.003 |
| | $\lambda_{(5–8)}$ | 0.800/left-skewed | **−0.950** | −0.018 | −0.003 |
| | $\lambda_{(9–12)}$ | 0.800/right-skewed | **−0.950** | −0.008 | −0.002 |
| | $\tau_{(1–4).1}$ | −1.645/normal | | −0.004 | 0.000 |
| | $\tau_{(1–4).2}$ | −0.643/normal | | −0.004 | −0.005 |
| | $\tau_{(1–4).3}$ | 0.643/normal | | −0.003 | −0.009 |
| | $\tau_{(1–4).4}$ | 1.645/normal | | 0.004 | 0.003 |
| | $\tau_{(5–8).1}$ | −1.622/left-skewed | | −0.006 | −0.002 |
| | $\tau_{(5–8).2}$ | −1.221/left-skewed | | −0.002 | −0.011 |
| | $\tau_{(5–8).3}$ | −0.800/left-skewed | | 0.003 | −0.009 |
| | $\tau_{(5–8).4}$ | −0.346/left-skewed | | 0.004 | −0.002 |
| | $\tau_{(9–12).1}$ | 0.346/right-skewed | | 0.005 | −0.006 |
| | $\tau_{(9–12).2}$ | 0.800/right-skewed | | 0.002 | −0.014 |
| | $\tau_{(9–12).3}$ | 1.221/right-skewed | | 0.001 | −0.011 |
| | $\tau_{(9–12).4}$ | 1.622/right-skewed | | 0.003 | −0.001 |
| bnNS6 | $\lambda_{(1–6)}$ | 0.800/normal | **−0.644** | −0.008 | 0.002 |
| | $\lambda_{(7–12)}$ | 0.800/bimodal | **−0.691** | −0.009 | −0.006 |
| | $\tau_{(1–6).1}$ | −1.645/normal | | 0.001 | −0.002 |
| | $\tau_{(1–6).2}$ | −0.643/normal | | 0.003 | −0.003 |
| | $\tau_{(1–6).3}$ | 0.643/normal | | 0.001 | −0.002 |
| | $\tau_{(1–6).4}$ | 1.645/normal | | 0.006 | 0.002 |
| | $\tau_{(7–12).1}$ | −1.282/bimodal | | 0.006 | −0.001 |
| | $\tau_{(7–12).2}$ | −0.126/bimodal | | −0.001 | 0.002 |
| | $\tau_{(7–12).3}$ | 0.126/bimodal | | −0.005 | −0.003 |
| | $\tau_{(7–12).4}$ | 1.282/bimodal | | 0.001 | −0.006 |

*Table E.63.* Coverage rate of 95%-confidence interval estimators for Cells nRS2, rnRS2, lnRS2, lrnRS2, and bnRS2 averaged over parameters with equal population value/shape. $R = 1000$ per parameter. Numbers represent deviation from 0.95, with values outside the range $(-0.05, 0.03)$ printed in boldface, indicating an unacceptable coverage rate.

| Cell | | $\omega$ | FA-lin | FA-poly | IRT-grm |
|---|---|---|---|---|---|
| nRS2 | $\lambda_{(1–12)}$ | 0.800/normal | **−0.372** | −0.029 | −0.001 |
| | $\tau_{(1–12).1}$ | −1.483/normal | | **−0.074** | 0.012 |
| | $\tau_{(1–12).2}$ | −0.680/normal | | −0.026 | 0.005 |
| | $\tau_{(1–12).3}$ | 0.583/normal | | −0.037 | 0.007 |
| | $\tau_{(1–12).4}$ | 1.790/normal | | **−0.144** | 0.005 |
| rnRS2 | $\lambda_{(1–6)}$ | 0.800/normal | **−0.576** | −0.035 | 0.002 |
| | $\lambda_{(7–12)}$ | 0.800/right-skewed | **−0.080** | **−0.457** | −0.027 |
| | $\tau_{(1–6).1}$ | −1.483/normal | | **−0.070** | 0.007 |
| | $\tau_{(1–6).2}$ | −0.680/normal | | −0.023 | −0.001 |
| | $\tau_{(1–6).3}$ | 0.583/normal | | −0.046 | 0.002 |
| | $\tau_{(1–6).4}$ | 1.790/normal | | **−0.149** | −0.006 |
| | $\tau_{(7–12).1}$ | 0.260/right-skewed | | **−0.109** | −0.004 |
| | $\tau_{(7–12).2}$ | 0.760/right-skewed | | −0.027 | −0.002 |
| | $\tau_{(7–12).3}$ | 1.260/right-skewed | | −0.015 | −0.003 |
| | $\tau_{(7–12).4}$ | 1.760/right-skewed | | **−0.111** | −0.003 |
| lnRS2 | $\lambda_{(1–6)}$ | 0.800/normal | **−0.424** | −0.016 | 0.002 |
| | $\lambda_{(7–12)}$ | 0.800/left-skewed | **−0.950** | **−0.529** | **−0.055** |
| | $\tau_{(1–6).1}$ | −1.483/normal | | **−0.079** | 0.009 |
| | $\tau_{(1–6).2}$ | −0.680/normal | | −0.020 | 0.006 |
| | $\tau_{(1–6).3}$ | 0.583/normal | | −0.032 | 0.003 |
| | $\tau_{(1–6).4}$ | 1.790/normal | | **−0.150** | 0.001 |
| | $\tau_{(7–12).1}$ | −1.465/left-skewed | | **−0.128** | 0.011 |
| | $\tau_{(7–12).2}$ | −1.156/left-skewed | | −0.014 | 0.005 |
| | $\tau_{(7–12).3}$ | −0.813/left-skewed | | 0.000 | 0.001 |
| | $\tau_{(7–12).4}$ | −0.416/left-skewed | | **−0.081** | −0.005 |
| lrnRS2 | $\lambda_{(1–4)}$ | 0.800/normal | **−0.365** | −0.030 | 0.000 |
| | $\lambda_{(5–8)}$ | 0.800/left-skewed | **−0.950** | **−0.497** | **−0.080** |
| | $\lambda_{(9–12)}$ | 0.800/right-skewed | **−0.107** | **−0.490** | −0.035 |
| | $\tau_{(1–4).1}$ | −1.483/normal | | **−0.076** | 0.004 |
| | $\tau_{(1–4).2}$ | −0.680/normal | | −0.031 | −0.004 |
| | $\tau_{(1–4).3}$ | 0.583/normal | | −0.040 | 0.002 |
| | $\tau_{(1–4).4}$ | 1.790/normal | | **−0.164** | −0.011 |
| | $\tau_{(5–8).1}$ | −1.465/left-skewed | | **−0.117** | 0.010 |
| | $\tau_{(5–8).2}$ | −1.156/left-skewed | | −0.010 | 0.013 |
| | $\tau_{(5–8).3}$ | −0.813/left-skewed | | 0.004 | 0.001 |
| | $\tau_{(5–8).4}$ | −0.416/left-skewed | | **−0.080** | −0.007 |
| | $\tau_{(9–12).1}$ | 0.260/right-skewed | | **−0.089** | 0.001 |
| | $\tau_{(9–12).2}$ | 0.760/right-skewed | | −0.013 | 0.004 |
| | $\tau_{(9–12).3}$ | 1.260/right-skewed | | −0.028 | −0.011 |
| | $\tau_{(9–12).4}$ | 1.760/right-skewed | | **−0.123** | −0.007 |
| bnRS2 | $\lambda_{(1–6)}$ | 0.800/normal | **−0.404** | −0.028 | −0.001 |
| | $\lambda_{(7–12)}$ | 0.800/bimodal | **−0.375** | −0.024 | −0.002 |
| | $\tau_{(1–6).1}$ | −1.483/normal | | **−0.078** | 0.010 |
| | $\tau_{(1–6).2}$ | −0.680/normal | | −0.026 | −0.002 |
| | $\tau_{(1–6).3}$ | 0.583/normal | | −0.037 | 0.003 |
| | $\tau_{(1–6).4}$ | 1.790/normal | | **−0.148** | 0.002 |
| | $\tau_{(7–12).1}$ | −1.203/bimodal | | −0.048 | −0.003 |
| | $\tau_{(7–12).2}$ | −0.212/bimodal | | **−0.128** | 0.000 |
| | $\tau_{(7–12).3}$ | 0.034/bimodal | | **−0.129** | 0.004 |
| | $\tau_{(7–12).4}$ | 1.334/bimodal | | −0.027 | 0.004 |

*Table E.64.* Coverage rate of 95%-confidence interval estimators for Cells nRS6, rnRS6, lnRS6, lrnRS6, and bnRS6 averaged over parameters with equal population value/shape. $R = 1000$ per parameter. Numbers represent deviation from 0.95, with values outside the range $(-0.05, 0.03)$ printed in boldface, indicating an unacceptable coverage rate.

| Cell | | $\omega$ | FA-lin | FA-poly | IRT-grm |
|---|---|---|---|---|---|
| nRS6 | $\lambda_{(1–12)}$ | 0.800/normal | **−0.869** | **−0.105** | −0.015 |
| | $\tau_{(1–12).1}$ | −1.483/normal | | **−0.434** | 0.001 |
| | $\tau_{(1–12).2}$ | −0.680/normal | | **−0.054** | −0.005 |
| | $\tau_{(1–12).3}$ | 0.583/normal | | **−0.143** | −0.006 |
| | $\tau_{(1–12).4}$ | 1.790/normal | | **−0.340** | −0.012 |
| rnRS6 | $\lambda_{(1–6)}$ | 0.800/normal | **−0.942** | **−0.060** | −0.028 |
| | $\lambda_{(7–12)}$ | 0.800/right-skewed | **−0.128** | **−0.772** | −0.020 |
| | $\tau_{(1–6).1}$ | −1.483/normal | | **−0.428** | −0.020 |
| | $\tau_{(1–6).2}$ | −0.680/normal | | **−0.051** | −0.013 |
| | $\tau_{(1–6).3}$ | 0.583/normal | | **−0.140** | 0.008 |
| | $\tau_{(1–6).4}$ | 1.790/normal | | **−0.327** | −0.016 |
| | $\tau_{(7–12).1}$ | 0.260/right-skewed | | **−0.340** | −0.004 |
| | $\tau_{(7–12).2}$ | 0.760/right-skewed | | **−0.055** | 0.009 |
| | $\tau_{(7–12).3}$ | 1.260/right-skewed | | −0.031 | −0.002 |
| | $\tau_{(7–12).4}$ | 1.760/right-skewed | | **−0.296** | −0.019 |
| lnRS6 | $\lambda_{(1–6)}$ | 0.800/normal | **−0.898** | **−0.087** | −0.021 |
| | $\lambda_{(7–12)}$ | 0.800/left-skewed | **−0.950** | **−0.936** | **−0.287** |
| | $\tau_{(1–6).1}$ | −1.483/normal | | **−0.418** | −0.003 |
| | $\tau_{(1–6).2}$ | −0.680/normal | | **−0.056** | 0.000 |
| | $\tau_{(1–6).3}$ | 0.583/normal | | **−0.134** | −0.003 |
| | $\tau_{(1–6).4}$ | 1.790/normal | | **−0.327** | −0.023 |
| | $\tau_{(7–12).1}$ | −1.465/left-skewed | | **−0.394** | −0.024 |
| | $\tau_{(7–12).2}$ | −1.156/left-skewed | | **−0.103** | 0.001 |
| | $\tau_{(7–12).3}$ | −0.813/left-skewed | | −0.009 | −0.013 |
| | $\tau_{(7–12).4}$ | −0.416/left-skewed | | **−0.232** | −0.036 |
| lrnRS6 | $\lambda_{(1–4)}$ | 0.800/normal | **−0.888** | **−0.087** | −0.028 |
| | $\lambda_{(5–8)}$ | 0.800/left-skewed | **−0.950** | **−0.931** | **−0.359** |
| | $\lambda_{(9–12)}$ | 0.800/right-skewed | **−0.312** | **−0.786** | −0.032 |
| | $\tau_{(1–4).1}$ | −1.483/normal | | **−0.434** | −0.025 |
| | $\tau_{(1–4).2}$ | −0.680/normal | | **−0.056** | −0.012 |
| | $\tau_{(1–4).3}$ | 0.583/normal | | **−0.148** | −0.006 |
| | $\tau_{(1–4).4}$ | 1.790/normal | | **−0.330** | −0.033 |
| | $\tau_{(5–8).1}$ | −1.465/left-skewed | | **−0.389** | −0.040 |
| | $\tau_{(5–8).2}$ | −1.156/left-skewed | | **−0.095** | −0.003 |
| | $\tau_{(5–8).3}$ | −0.813/left-skewed | | −0.014 | −0.021 |
| | $\tau_{(5–8).4}$ | −0.416/left-skewed | | **−0.236** | **−0.066** |
| | $\tau_{(9–12).1}$ | 0.260/right-skewed | | **−0.334** | −0.018 |
| | $\tau_{(9–12).2}$ | 0.760/right-skewed | | **−0.061** | −0.002 |
| | $\tau_{(9–12).3}$ | 1.260/right-skewed | | −0.050 | −0.021 |
| | $\tau_{(9–12).4}$ | 1.760/right-skewed | | **−0.308** | −0.031 |
| bnRS6 | $\lambda_{(1–6)}$ | 0.800/normal | **−0.890** | **−0.099** | −0.011 |
| | $\lambda_{(7–12)}$ | 0.800/bimodal | **−0.871** | **−0.089** | −0.043 |
| | $\tau_{(1–6).1}$ | −1.483/normal | | **−0.429** | −0.002 |
| | $\tau_{(1–6).2}$ | −0.680/normal | | **−0.056** | −0.005 |
| | $\tau_{(1–6).3}$ | 0.583/normal | | **−0.141** | 0.006 |
| | $\tau_{(1–6).4}$ | 1.790/normal | | **−0.321** | 0.000 |
| | $\tau_{(7–12).1}$ | −1.203/bimodal | | **−0.146** | −0.001 |
| | $\tau_{(7–12).2}$ | −0.212/bimodal | | **−0.332** | −0.016 |
| | $\tau_{(7–12).3}$ | 0.034/bimodal | | **−0.369** | −0.009 |
| | $\tau_{(7–12).4}$ | 1.334/bimodal | | **−0.071** | 0.001 |

### E.1.3 Average Loevinger's $H$ Results for IRT-mok

*Table E.65.* IRT-mok $H_i$ results for Cell nNS2 averaged over parameters with equal population value. $n = 200$; $R = 1000$.

| | $\omega$ | $\bar{\hat{\omega}}$ | PB($\hat{\omega}$) | RB($\hat{\omega}$) | RMSE($\hat{\omega}$) | SD($\hat{\omega}$) | RB($\hat{se}$) | RMSE($\hat{se}$) | 95%-cov. |
|---|---|---|---|---|---|---|---|---|---|
| $H_{(1–12)}$ | 0.571 | 0.599 | 0.027 | 0.048 | 0.043 | 0.033 | 0.027 | 0.003 | 0.859 |
| $H_{scale}$ | 0.571 | 0.599 | 0.027 | 0.048 | 0.038 | 0.026 | 0.030 | 0.002 | 0.827 |

*Table E.66.* IRT-mok $H_i$ results for Cell nNS6 averaged over parameters with equal population value. $n = 600$; $R = 1000$.

| | $\omega$ | $\bar{\hat{\omega}}$ | PB($\hat{\omega}$) | RB($\hat{\omega}$) | RMSE($\hat{\omega}$) | SD($\hat{\omega}$) | RB($\hat{se}$) | RMSE($\hat{se}$) | 95%-cov. |
|---|---|---|---|---|---|---|---|---|---|
| $H_{(1–12)}$ | 0.571 | 0.588 | 0.017 | 0.029 | 0.026 | 0.019 | 0.021 | 0.001 | 0.856 |
| $H_{scale}$ | 0.571 | 0.588 | 0.017 | 0.029 | 0.023 | 0.016 | 0.006 | 0.001 | 0.799 |

*Table E.67.* IRT-mok $H_i$ results for Cell rnNS2 averaged over parameters with equal population value. $n = 200$; $R = 1000$.

| | $\omega$ | $\bar{\hat{\omega}}$ | PB($\hat{\omega}$) | RB($\hat{\omega}$) | RMSE($\hat{\omega}$) | SD($\hat{\omega}$) | RB($\hat{se}$) | RMSE($\hat{se}$) | 95%-cov. |
|---|---|---|---|---|---|---|---|---|---|
| $H_{(1–6)}$ | 0.604 | 0.619 | 0.015 | 0.025 | 0.038 | 0.035 | 0.010 | 0.003 | 0.916 |
| $H_{(7–12)}$ | 0.572 | 0.583 | 0.010 | 0.018 | 0.044 | 0.043 | −0.005 | 0.004 | 0.930 |
| $H_{scale}$ | 0.586 | 0.599 | 0.013 | 0.021 | 0.033 | 0.031 | 0.003 | 0.003 | 0.918 |

*Table E.68.* IRT-mok $H_i$ results for Cell rnNS6 averaged over parameters with equal population value. $n = 600$; $R = 1000$.

| | $\omega$ | $\bar{\hat{\omega}}$ | PB($\hat{\omega}$) | RB($\hat{\omega}$) | RMSE($\hat{\omega}$) | SD($\hat{\omega}$) | RB($\hat{se}$) | RMSE($\hat{se}$) | 95%-cov. |
|---|---|---|---|---|---|---|---|---|---|
| $H_{(1–6)}$ | 0.604 | 0.613 | 0.009 | 0.014 | 0.023 | 0.021 | −0.021 | 0.001 | 0.913 |
| $H_{(7–12)}$ | 0.572 | 0.579 | 0.007 | 0.011 | 0.026 | 0.026 | −0.031 | 0.002 | 0.927 |
| $H_{scale}$ | 0.586 | 0.594 | 0.008 | 0.013 | 0.020 | 0.019 | −0.042 | 0.001 | 0.914 |

*Table E.69.* IRT-mok $H_i$ results for Cell lnNS2 averaged over parameters with equal population value. $n = 200$; $R = 1000$.

| | $\omega$ | $\bar{\hat{\omega}}$ | PB($\hat{\omega}$) | RB($\hat{\omega}$) | RMSE($\hat{\omega}$) | SD($\hat{\omega}$) | RB($\hat{se}$) | RMSE($\hat{se}$) | 95%-cov. |
|---|---|---|---|---|---|---|---|---|---|
| $H_{(1–6)}$ | 0.604 | 0.617 | 0.012 | 0.020 | 0.037 | 0.035 | 0.016 | 0.003 | 0.927 |
| $H_{(7–12)}$ | 0.572 | 0.580 | 0.007 | 0.013 | 0.043 | 0.043 | 0.006 | 0.004 | 0.936 |
| $H_{scale}$ | 0.586 | 0.596 | 0.010 | 0.017 | 0.032 | 0.030 | 0.024 | 0.003 | 0.943 |

*Table E.70.* IRT-mok $H_i$ results for Cell rnNS6 averaged over parameters with equal population value. $n = 600$; $R = 1000$.

| | $\omega$ | $\bar{\hat{\omega}}$ | PB($\hat{\omega}$) | RB($\hat{\omega}$) | RMSE($\hat{\omega}$) | SD($\hat{\omega}$) | RB($\hat{se}$) | RMSE($\hat{se}$) | 95%-cov. |
|---|---|---|---|---|---|---|---|---|---|
| $H_{(1–6)}$ | 0.604 | 0.613 | 0.009 | 0.014 | 0.023 | 0.021 | −0.021 | 0.001 | 0.913 |
| $H_{(7–12)}$ | 0.572 | 0.579 | 0.007 | 0.011 | 0.026 | 0.026 | −0.031 | 0.002 | 0.927 |
| $H_{scale}$ | 0.586 | 0.594 | 0.008 | 0.013 | 0.020 | 0.019 | −0.042 | 0.001 | 0.914 |

*Table E.71.* IRT-mok $H_i$ results for Cell lrnNS2 averaged over parameters with equal population value. $n = 200$; $R = 1000$.

| | $\omega$ | $\bar{\hat{\omega}}$ | PB($\hat{\omega}$) | RB($\hat{\omega}$) | RMSE($\hat{\omega}$) | SD($\hat{\omega}$) | RB($\hat{se}$) | RMSE($\hat{se}$) | 95%-cov. |
|---|---|---|---|---|---|---|---|---|---|
| $H_{(1–4)}$ | 0.615 | 0.624 | 0.008 | 0.014 | 0.035 | 0.034 | −0.003 | 0.003 | 0.927 |
| $H_{(5–12)}$ | 0.635 | 0.641 | 0.006 | 0.010 | 0.041 | 0.041 | −0.030 | 0.005 | 0.930 |
| $H_{scale}$ | 0.629 | 0.636 | 0.007 | 0.012 | 0.028 | 0.027 | −0.012 | 0.002 | 0.923 |

*Table E.72.* IRT-mok $H_i$ results for Cell lrnNS6 averaged over parameters with equal population value. $n = 600$; $R = 1000$.

| | $\omega$ | $\bar{\hat{\omega}}$ | PB($\hat{\omega}$) | RB($\hat{\omega}$) | RMSE($\hat{\omega}$) | SD($\hat{\omega}$) | RB($\hat{se}$) | RMSE($\hat{se}$) | 95%-cov. |
|---|---|---|---|---|---|---|---|---|---|
| $H_{(1–4)}$ | 0.615 | 0.621 | 0.006 | 0.010 | 0.020 | 0.019 | 0.019 | 0.001 | 0.939 |
| $H_{(5–12)}$ | 0.635 | 0.641 | 0.006 | 0.009 | 0.023 | 0.022 | 0.011 | 0.001 | 0.937 |
| $H_{scale}$ | 0.629 | 0.634 | 0.006 | 0.009 | 0.016 | 0.015 | 0.025 | 0.001 | 0.934 |

*Table E.73.* IRT-mok $H_i$ results for Cell bnNS2 averaged over parameters with equal population value. $n = 200$; $R = 1000$.

| | $\omega$ | $\bar{\hat{\omega}}$ | PB($\hat{\omega}$) | RB($\hat{\omega}$) | RMSE($\hat{\omega}$) | SD($\hat{\omega}$) | RB($\hat{se}$) | RMSE($\hat{se}$) | 95%-cov. |
|---|---|---|---|---|---|---|---|---|---|
| $H_{(1–6)}$ | 0.633 | 0.645 | 0.011 | 0.018 | 0.035 | 0.033 | 0.005 | 0.003 | 0.919 |
| $H_{(7–12)}$ | 0.614 | 0.624 | 0.010 | 0.016 | 0.036 | 0.034 | −0.004 | 0.003 | 0.924 |
| $H_{scale}$ | 0.623 | 0.633 | 0.011 | 0.017 | 0.028 | 0.026 | 0.004 | 0.002 | 0.919 |

*Table E.74.* IRT-mok $H_i$ results for Cell bnNS6 averaged over parameters with equal population value. $n = 600$; $R = 1000$.

| | $\omega$ | $\bar{\hat{\omega}}$ | PB($\hat{\omega}$) | RB($\hat{\omega}$) | RMSE($\hat{\omega}$) | SD($\hat{\omega}$) | RB($\hat{se}$) | RMSE($\hat{se}$) | 95%-cov. |
|---|---|---|---|---|---|---|---|---|---|
| $H_{(1–6)}$ | 0.633 | 0.641 | 0.008 | 0.012 | 0.020 | 0.019 | 0.015 | 0.001 | 0.921 |
| $H_{(7–12)}$ | 0.614 | 0.621 | 0.007 | 0.011 | 0.021 | 0.019 | 0.017 | 0.001 | 0.933 |
| $H_{scale}$ | 0.623 | 0.630 | 0.007 | 0.011 | 0.016 | 0.015 | 0.027 | 0.001 | 0.907 |

*Table E.75.* IRT-mok $H_i$ results for Cell nRS2 averaged over parameters with equal population value. $n = 200$; $R = 1000$.

| | $\omega$ | $\bar{\hat{\omega}}$ | PB($\hat{\omega}$) | RB($\hat{\omega}$) | RMSE($\hat{\omega}$) | SD($\hat{\omega}$) | RB($\hat{se}$) | RMSE($\hat{se}$) | 95%-cov. |
|---|---|---|---|---|---|---|---|---|---|
| $H_{(1–12)}$ | 0.542 | 0.568 | 0.026 | 0.048 | 0.043 | 0.034 | 0.028 | 0.003 | 0.872 |
| $H_{scale}$ | 0.542 | 0.568 | 0.026 | 0.048 | 0.038 | 0.028 | 0.032 | 0.002 | 0.844 |

*Table E.76.* IRT-mok $H_i$ results for Cell nRS6 averaged over parameters with equal population value. $n = 600$; $R = 1000$.

| | $\omega$ | $\bar{\hat{\omega}}$ | PB($\hat{\omega}$) | RB($\hat{\omega}$) | RMSE($\hat{\omega}$) | SD($\hat{\omega}$) | RB($\hat{se}$) | RMSE($\hat{se}$) | 95%-cov. |
|---|---|---|---|---|---|---|---|---|---|
| $H_{(1–12)}$ | 0.542 | 0.558 | 0.016 | 0.029 | 0.026 | 0.021 | −0.010 | 0.001 | 0.862 |
| $H_{scale}$ | 0.542 | 0.558 | 0.016 | 0.029 | 0.023 | 0.017 | −0.029 | 0.001 | 0.827 |

*Table E.77.* IRT-mok $H_i$ results for Cell rnRS2 averaged over parameters with equal population value. $n = 200$; $R = 1000$.

| | $\omega$ | $\bar{\hat{\omega}}$ | PB($\hat{\omega}$) | RB($\hat{\omega}$) | RMSE($\hat{\omega}$) | SD($\hat{\omega}$) | RB($\hat{se}$) | RMSE($\hat{se}$) | 95%-cov. |
|---|---|---|---|---|---|---|---|---|---|
| $H_{(1–6)}$ | 0.630 | 0.643 | 0.013 | 0.021 | 0.038 | 0.036 | −0.002 | 0.004 | 0.917 |
| $H_{(7–12)}$ | 0.660 | 0.669 | 0.008 | 0.013 | 0.042 | 0.041 | −0.015 | 0.005 | 0.929 |
| $H_{scale}$ | 0.647 | 0.658 | 0.011 | 0.016 | 0.034 | 0.032 | −0.014 | 0.003 | 0.930 |

*Table E.78.* IRT-mok $H_i$ results for Cell rnRS6 averaged over parameters with equal population value. $n = 600$; $R = 1000$.

| | $\omega$ | $\bar{\hat{\omega}}$ | PB($\hat{\omega}$) | RB($\hat{\omega}$) | RMSE($\hat{\omega}$) | SD($\hat{\omega}$) | RB($\hat{se}$) | RMSE($\hat{se}$) | 95%-cov. |
|---|---|---|---|---|---|---|---|---|---|
| $H_{(1–6)}$ | 0.630 | 0.638 | 0.008 | 0.013 | 0.022 | 0.021 | 0.003 | 0.001 | 0.921 |
| $H_{(7–12)}$ | 0.660 | 0.667 | 0.006 | 0.009 | 0.024 | 0.023 | 0.011 | 0.002 | 0.935 |
| $H_{scale}$ | 0.647 | 0.654 | 0.007 | 0.011 | 0.019 | 0.018 | 0.013 | 0.001 | 0.923 |

*Table E.79.* IRT-mok $H_i$ results for Cell lnRS2 averaged over parameters with equal population value. $n = 200$; $R = 1000$.

| | $\omega$ | $\bar{\hat{\omega}}$ | PB($\hat{\omega}$) | RB($\hat{\omega}$) | RMSE($\hat{\omega}$) | SD($\hat{\omega}$) | RB($\hat{se}$) | RMSE($\hat{se}$) | 95%-cov. |
|---|---|---|---|---|---|---|---|---|---|
| $H_{(1–6)}$ | 0.517 | 0.531 | 0.014 | 0.027 | 0.039 | 0.036 | −0.015 | 0.003 | 0.914 |
| $H_{(7–12)}$ | 0.417 | 0.426 | 0.010 | 0.023 | 0.041 | 0.040 | −0.020 | 0.003 | 0.933 |
| $H_{scale}$ | 0.460 | 0.472 | 0.012 | 0.026 | 0.030 | 0.027 | −0.029 | 0.002 | 0.920 |

*Table E.80.* IRT-mok $H_i$ results for Cell lnRS6 averaged over parameters with equal population value. $n = 600$; $R = 1000$.

| | $\omega$ | $\bar{\hat{\omega}}$ | PB($\hat{\omega}$) | RB($\hat{\omega}$) | RMSE($\hat{\omega}$) | SD($\hat{\omega}$) | RB($\hat{se}$) | RMSE($\hat{se}$) | 95%-cov. |
|---|---|---|---|---|---|---|---|---|---|
| $H_{(1–6)}$ | 0.517 | 0.524 | 0.007 | 0.014 | 0.022 | 0.020 | 0.008 | 0.001 | 0.934 |
| $H_{(7–12)}$ | 0.417 | 0.422 | 0.005 | 0.012 | 0.023 | 0.022 | 0.008 | 0.001 | 0.944 |
| $H_{scale}$ | 0.460 | 0.466 | 0.006 | 0.013 | 0.016 | 0.015 | 0.019 | 0.001 | 0.936 |

*Table E.81.* IRT-mok $H_i$ results for Cell lrnRS2 averaged over parameters with equal population value. $n = 200$; $R = 1000$.

| | $\omega$ | $\bar{\hat{\omega}}$ | PB($\hat{\omega}$) | RB($\hat{\omega}$) | RMSE($\hat{\omega}$) | SD($\hat{\omega}$) | RB($\hat{se}$) | RMSE($\hat{se}$) | 95%-cov. |
|---|---|---|---|---|---|---|---|---|---|
| $H_{(1–4)}$ | 0.584 | 0.594 | 0.010 | 0.018 | 0.037 | 0.036 | −0.024 | 0.003 | 0.917 |
| $H_{(5–8)}$ | 0.512 | 0.521 | 0.009 | 0.017 | 0.042 | 0.041 | −0.030 | 0.004 | 0.929 |
| $H_{(9–12)}$ | 0.699 | 0.708 | 0.009 | 0.013 | 0.038 | 0.037 | −0.015 | 0.005 | 0.922 |
| $H_{scale}$ | 0.599 | 0.608 | 0.010 | 0.016 | 0.030 | 0.029 | −0.030 | 0.002 | 0.916 |

*Table E.82.* IRT-mok $H_i$ results for Cell lrnRS6 averaged over parameters with equal population value. $n = 600$; $R = 1000$.

| | $\omega$ | $\bar{\hat{\omega}}$ | PB($\hat{\omega}$) | RB($\hat{\omega}$) | RMSE($\hat{\omega}$) | SD($\hat{\omega}$) | RB($\hat{se}$) | RMSE($\hat{se}$) | 95%-cov. |
|---|---|---|---|---|---|---|---|---|---|
| $H_{(1–4)}$ | 0.584 | 0.590 | 0.006 | 0.010 | 0.021 | 0.021 | −0.013 | 0.001 | 0.934 |
| $H_{(5–8)}$ | 0.512 | 0.516 | 0.004 | 0.008 | 0.024 | 0.023 | −0.013 | 0.001 | 0.942 |
| $H_{(9–12)}$ | 0.699 | 0.705 | 0.005 | 0.008 | 0.022 | 0.022 | −0.011 | 0.002 | 0.930 |
| $H_{scale}$ | 0.599 | 0.604 | 0.005 | 0.009 | 0.017 | 0.017 | −0.025 | 0.001 | 0.921 |

*Table E.83.* IRT-mok $H_i$ results for Cell bnRS2 averaged over parameters with equal population value. $n = 200$; $R = 1000$.

| | $\omega$ | $\bar{\hat{\omega}}$ | PB($\hat{\omega}$) | RB($\hat{\omega}$) | RMSE($\hat{\omega}$) | SD($\hat{\omega}$) | RB($\hat{se}$) | RMSE($\hat{se}$) | 95%-cov. |
|---|---|---|---|---|---|---|---|---|---|
| $H_{(1–6)}$ | 0.603 | 0.613 | 0.010 | 0.016 | 0.036 | 0.034 | −0.006 | 0.004 | 0.923 |
| $H_{(7–12)}$ | 0.588 | 0.597 | 0.008 | 0.014 | 0.036 | 0.035 | 0.004 | 0.003 | 0.935 |
| $H_{scale}$ | 0.595 | 0.603 | 0.009 | 0.015 | 0.029 | 0.027 | 0.002 | 0.002 | 0.930 |

*Table E.84.* IRT-mok $H_i$ results for Cell bnRS6 averaged over parameters with equal population value. $n = 600$; $R = 1000$.

| | $\omega$ | $\bar{\hat{\omega}}$ | PB($\hat{\omega}$) | RB($\hat{\omega}$) | RMSE($\hat{\omega}$) | SD($\hat{\omega}$) | RB($\hat{se}$) | RMSE($\hat{se}$) | 95%-cov. |
|---|---|---|---|---|---|---|---|---|---|
| $H_{(1–6)}$ | 0.603 | 0.610 | 0.007 | 0.012 | 0.021 | 0.020 | 0.005 | 0.001 | 0.927 |
| $H_{(7–12)}$ | 0.588 | 0.595 | 0.006 | 0.010 | 0.021 | 0.020 | −0.005 | 0.001 | 0.932 |
| $H_{scale}$ | 0.595 | 0.601 | 0.007 | 0.011 | 0.017 | 0.016 | −0.009 | 0.001 | 0.925 |

## E.2 Step-Difficulty Parameter Estimation Results

*Table E.85.* MANOVA results: $\eta_p^2$ per effect for PB of $\beta$ parameters. $N = 40000$.

| Effect | Levels | $\eta_p^2$ for PB of parameters |
|---|---|---|
| Model (m) | 2 | 0.065 |
| LV distribution (lv) | 2 | 0.152 |
| Scale shape (ss) | 5 | 0.067 |
| Sample size (n) | 2 | |
| m × lv | 4 | 0.065 |
| lv × ss | 10 | 0.059 |
| Item group (ig) | 6 | 0.083 |
| lv × ig | 12 | 0.075 |
| ss × ig | 30 | 0.099 |
| lv × ss × ig | 60 | 0.083 |
| Threshold type (t) | 2 | 0.664 |
| m × t | 4 | 0.499 |
| lv × t | 4 | 0.663 |
| ss × t | 10 | 0.145 |
| m × lv × t | 8 | 0.498 |
| m × ss × t | 20 | 0.147 |
| lv × ss × t | 20 | 0.139 |
| m × lv × ss × t | 40 | 0.145 |
| ig × t | 12 | 0.351 |
| m × ig × t | 24 | 0.157 |
| lv × ig × t | 24 | 0.342 |
| ss × ig × t | 60 | 0.071 |
| m × lv × ig × t | 48 | 0.163 |
| m × ss × ig × t | 120 | 0.037 |
| lv × ss × ig × t | 120 | 0.074 |
| m × lv × ss × ig × t | 240 | 0.031 |

*Note.* Listed effects are statistically significant at $\alpha = 0.01$ and are sized $\eta_p^2 > 0.02$.

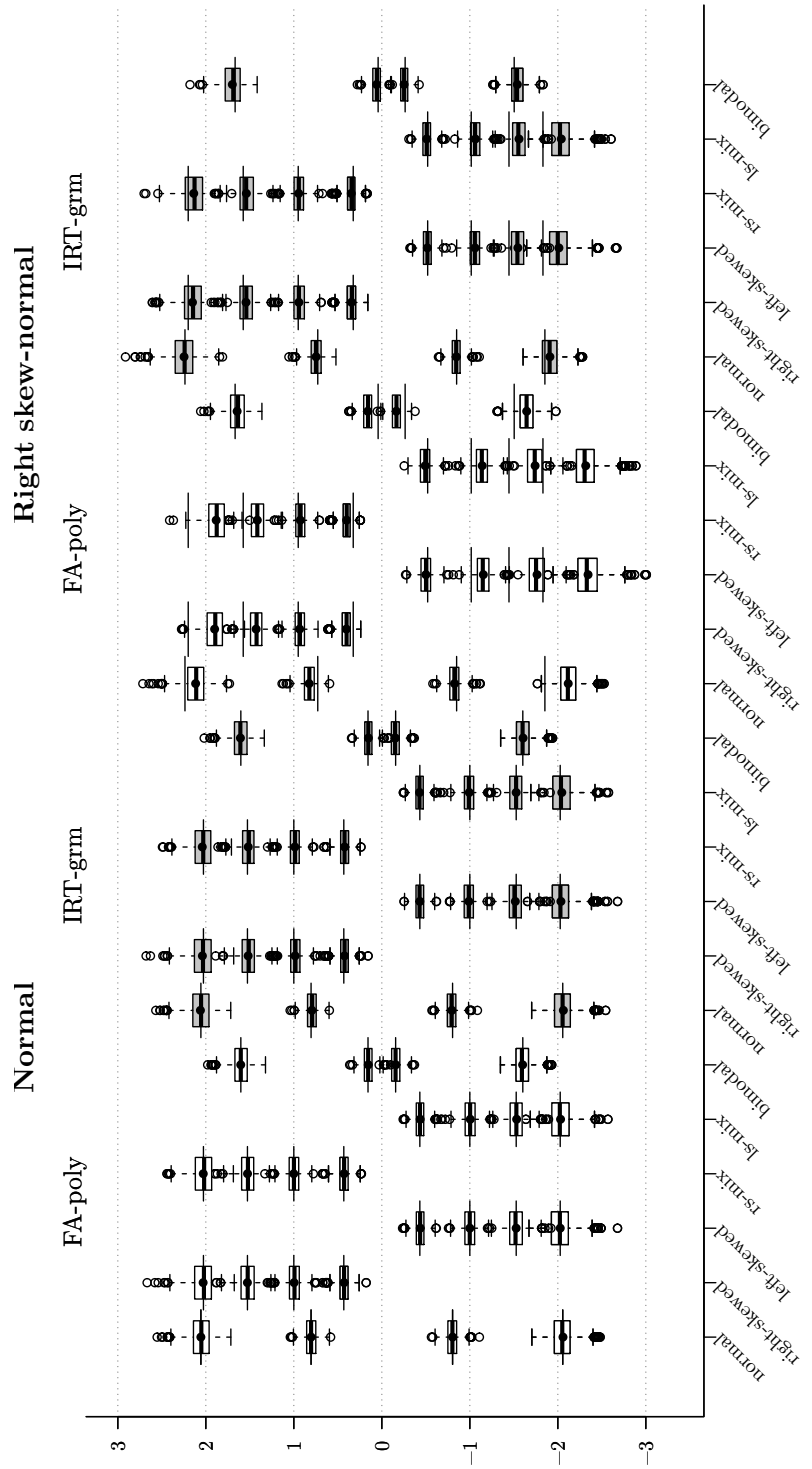



*Figure E.1.* Parameter estimates $\hat{\beta}_{ic}$ for a normal and right skew-normal LV and variously shaped items, as estimated by FA-poly and IRT-grm. The horizontal lines crossing the boxplots represent the true values. $n =$; $R = 1000$.

## E.3 Additional Fit Results: RMSEA for medium sample size

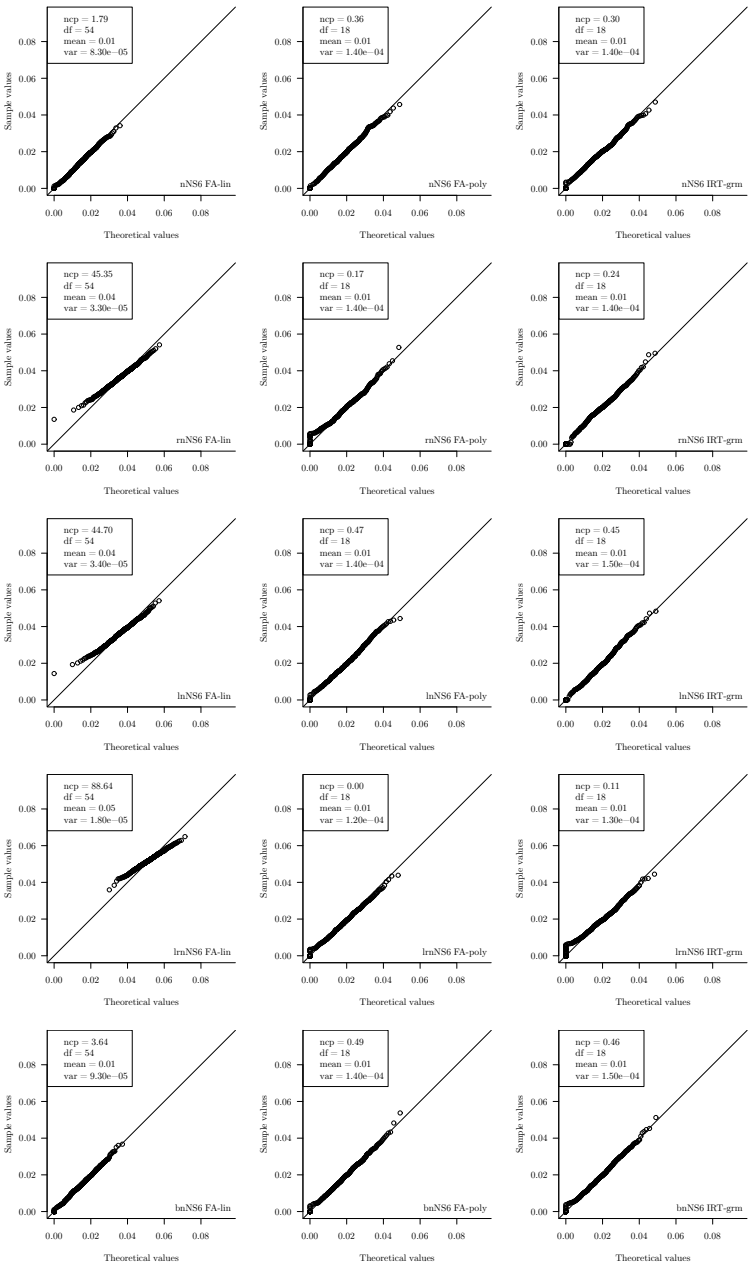



*Figure E.2.* Q-Q plots for RMSEA fit statistic for Cells nNS6, rnNS6, lnNS6, lrnNS6, and bnNS6 and each model. LV distribution is normal. $n = 600$; $R = 1000$. The diagonal line depicts a perfect association between the empirical and theoretical distribution, the latter being a noncentral $\chi^2$ distribution using the mean empirical noncentrality parameter (NCP) over $R$ replications.

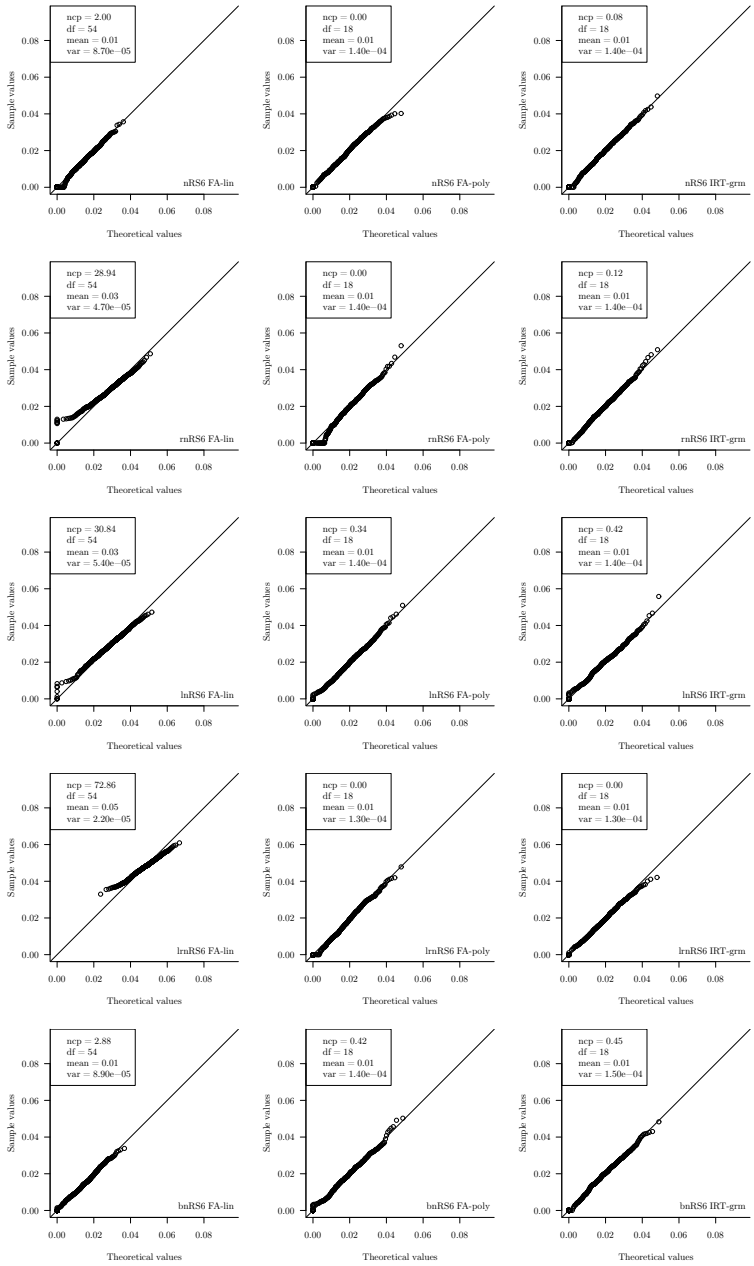



*Figure E.3.* Q-Q plots for RMSEA fit statistic for Cells nRS6, rnRS6, lnRS6, lrnRS6, and bnRS6 and each model. LV distribution is right skew-normal. $n = 600$; $R = 1000$. The diagonal line depicts a perfect association between the empirical and theoretical distribution, the latter being a noncentral $\chi^2$ distribution using the mean empirical noncentrality parameter (NCP) over $R$ replications.

## E.4 Additional LV Results: medium sample size

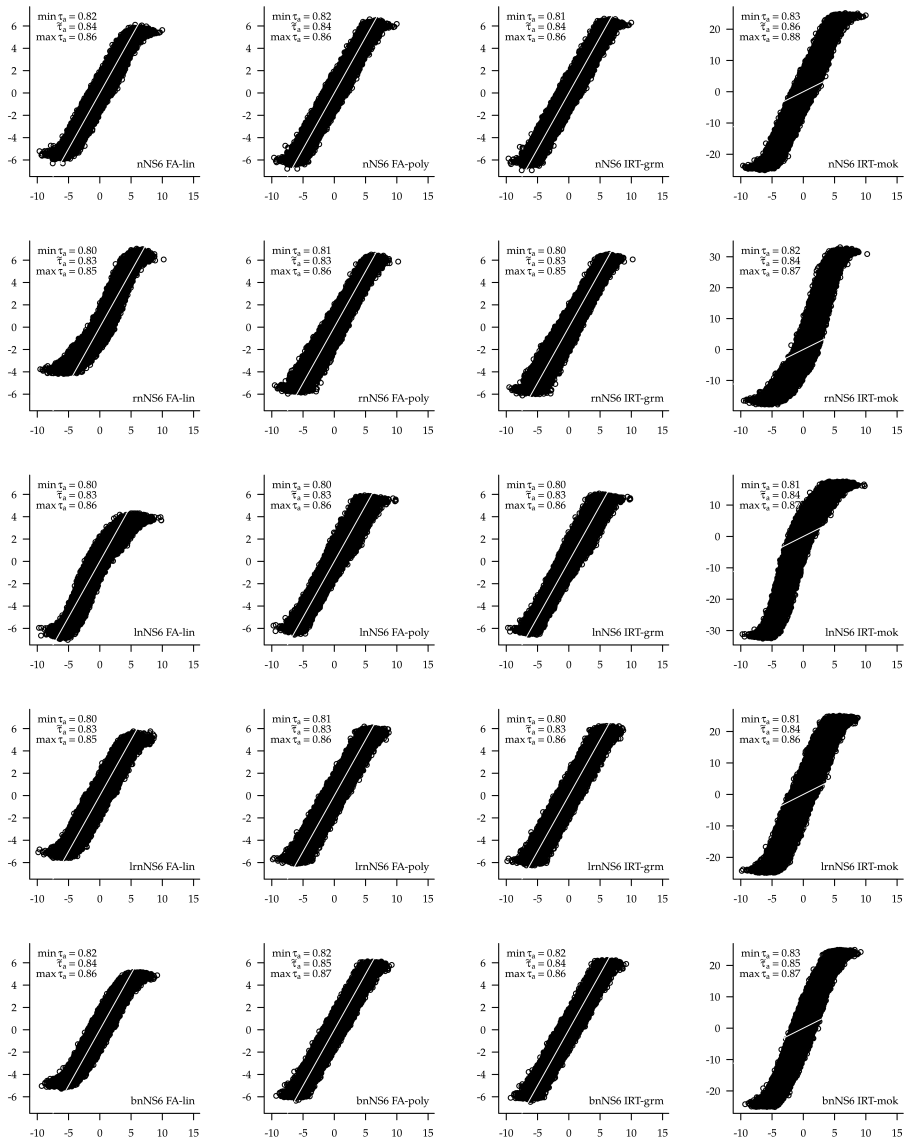



*Figure E.4.* Scatterplots of LV population $\theta_{sr} - \bar{\theta}_r$ ($x$-axis) and estimated $\hat{\theta}_{sr} - \bar{\hat{\theta}}_r$ ($y$-axis) deviation scores for FA-lin, FA-poly, IRT-grm, and IRT-mok in Cells nNS6, rnNS6, lnNS6, lrnNS6, and bnNS6 for each replication. LV distribution is normal. $n = 600$; $R = 1000$. The minimum, median, and maximum Kendall's $\tau_a$ over replications are given in the inset of each plot as an indication of association.

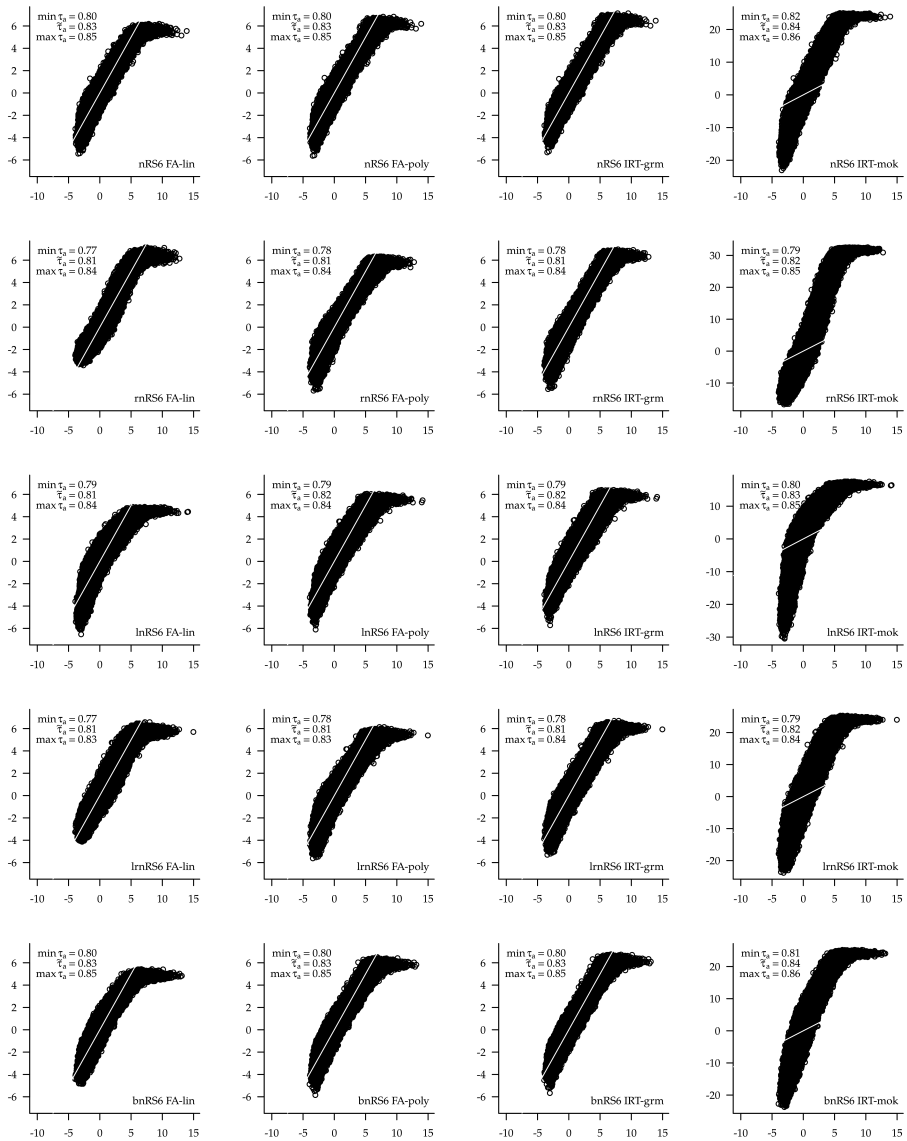



*Figure E.5.* Scatterplots of LV population $\theta_{sr} - \bar{\theta}_r$ ($x$-axis) and estimated $\hat{\theta}_{sr} - \bar{\hat{\theta}}_r$ ($y$-axis) deviation scores for FA-lin, FA-poly, IRT-grm, and IRT-mok in Cells nRS6, rnRS6, lnRS6, lrnRS6, and bnRS6 for each replication. LV distribution is right skew-normal. $n = 600$; $R = 1000$. The minimum, median, and maximum Kendall's $\tau_a$ over replications are given in the inset of each plot as an indication of association.

# Appendix F

# Applications of FA and IRT

## F.1 Additional Results for DBIQ

### F.1.1 Threshold Parameter Estimation Results

*Table F.1.* Threshold parameter and standard error estimates for DBIQ-*Body acceptance.* $n = 761$.

| | FA-poly | | IRT-grm | |
|---|---|---|---|---|
| | $\hat{\tau}$ | $\hat{se}(\hat{\tau})$ | $\hat{\tau}$ | $\hat{se}(\hat{\tau})$ |
| $\tau_{1.1}$ | −2.414 | 0.148 | −2.541 | 0.163 |
| $\tau_{1.2}$ | −1.528 | 0.071 | −1.553 | 0.074 |
| $\tau_{1.3}$ | −0.609 | 0.049 | −0.583 | 0.049 |
| $\tau_{1.4}$ | 0.981 | 0.054 | 0.953 | 0.052 |
| $\tau_{2.1}$ | −2.059 | 0.105 | −2.148 | 0.112 |
| $\tau_{2.2}$ | −1.288 | 0.062 | −1.288 | 0.064 |
| $\tau_{2.3}$ | −0.190 | 0.046 | −0.160 | 0.045 |
| $\tau_{2.4}$ | 1.280 | 0.062 | 1.244 | 0.060 |
| $\tau_{3.1}$ | −2.412 | 0.148 | −2.564 | 0.190 |
| $\tau_{3.2}$ | −1.272 | 0.062 | −1.252 | 0.063 |
| $\tau_{3.3}$ | −0.472 | 0.047 | −0.447 | 0.045 |
| $\tau_{3.4}$ | 0.491 | 0.048 | 0.465 | 0.045 |
| $\tau_{4.1}$ | −2.789 | 0.229 | −2.937 | 0.257 |
| $\tau_{4.2}$ | −1.456 | 0.068 | −1.467 | 0.070 |
| $\tau_{4.3}$ | −0.775 | 0.051 | −0.755 | 0.050 |
| $\tau_{4.4}$ | 0.378 | 0.047 | 0.385 | 0.044 |
| $\tau_{5.1}$ | −2.084 | 0.108 | −2.160 | 0.112 |
| $\tau_{5.2}$ | −1.465 | 0.069 | −1.478 | 0.071 |
| $\tau_{5.3}$ | −0.711 | 0.050 | −0.676 | 0.049 |
| $\tau_{5.4}$ | 0.065 | 0.046 | 0.097 | 0.044 |
| $\tau_{6.1}$ | −2.480 | 0.159 | −2.564 | 0.170 |
| $\tau_{6.2}$ | −1.585 | 0.074 | −1.595 | 0.077 |
| $\tau_{6.3}$ | −0.492 | 0.047 | −0.469 | 0.047 |
| $\tau_{6.4}$ | 1.089 | 0.057 | 1.067 | 0.055 |
| $\tau_{7.1}$ | −1.713 | 0.080 | −1.718 | 0.089 |
| $\tau_{7.2}$ | −0.769 | 0.051 | −0.733 | 0.049 |
| $\tau_{7.3}$ | −0.196 | 0.046 | −0.182 | 0.043 |
| $\tau_{7.4}$ | 0.586 | 0.048 | 0.554 | 0.045 |
| $\tau_{8.1}$ | −1.412 | 0.066 | −1.390 | 0.071 |
| $\tau_{8.2}$ | −0.425 | 0.047 | −0.400 | 0.044 |
| $\tau_{8.3}$ | 0.755 | 0.051 | 0.696 | 0.048 |
| $\tau_{8.4}$ | 2.088 | 0.108 | 2.117 | 0.134 |

*Table F.2.* Threshold parameter and standard error estimates for DBIQ-*Sexual fulfillment*. $n = 761$.

| | FA-poly | | IRT-grm | |
|---|---|---|---|---|
| | $\hat{\tau}$ | $\hat{se}(\hat{\tau})$ | $\hat{\tau}$ | $\hat{se}(\hat{\tau})$ |
| $\tau_{1.1}$ | −1.957 | 0.097 | −2.084 | 0.112 |
| $\tau_{1.2}$ | −1.536 | 0.072 | −1.560 | 0.074 |
| $\tau_{1.3}$ | −0.678 | 0.050 | −0.638 | 0.049 |
| $\tau_{1.4}$ | 0.682 | 0.050 | 0.654 | 0.047 |
| $\tau_{2.1}$ | −1.956 | 0.097 | −2.070 | 0.104 |
| $\tau_{2.2}$ | −1.373 | 0.065 | −1.355 | 0.070 |
| $\tau_{2.3}$ | −0.495 | 0.048 | −0.460 | 0.047 |
| $\tau_{2.4}$ | 0.767 | 0.051 | 0.746 | 0.049 |
| $\tau_{3.1}$ | −1.916 | 0.094 | −1.974 | 0.106 |
| $\tau_{3.2}$ | −1.030 | 0.055 | −0.995 | 0.056 |
| $\tau_{3.3}$ | −0.015 | 0.045 | −0.013 | 0.043 |
| $\tau_{3.4}$ | 1.094 | 0.057 | 1.046 | 0.055 |
| $\tau_{4.1}$ | −1.915 | 0.094 | −2.004 | 0.106 |
| $\tau_{4.2}$ | −1.173 | 0.059 | −1.154 | 0.061 |
| $\tau_{4.3}$ | −0.279 | 0.046 | −0.270 | 0.044 |
| $\tau_{4.4}$ | 0.805 | 0.051 | 0.771 | 0.049 |
| $\tau_{5.1}$ | −2.112 | 0.111 | −2.325 | 0.134 |
| $\tau_{5.2}$ | −1.500 | 0.070 | −1.507 | 0.076 |
| $\tau_{5.3}$ | −0.671 | 0.050 | −0.630 | 0.050 |
| $\tau_{5.4}$ | 0.650 | 0.049 | 0.614 | 0.047 |
| $\tau_{6.1}$ | −1.914 | 0.094 | −2.011 | 0.103 |
| $\tau_{6.2}$ | −1.301 | 0.063 | −1.271 | 0.066 |
| $\tau_{6.3}$ | −0.500 | 0.048 | −0.459 | 0.047 |
| $\tau_{6.4}$ | 0.827 | 0.052 | 0.800 | 0.050 |

*Table F.3.* Threshold parameter and standard error estimates for DBIQ-*Self-aggrandizement*. $n = 761$.

| | FA-poly | | IRT-grm | |
|---|---|---|---|---|
| | $\hat{\tau}$ | $\hat{se}(\hat{\tau})$ | $\hat{\tau}$ | $\hat{se}(\hat{\tau})$ |
| $\tau_{1.1}$ | −1.876 | 0.091 | −1.899 | 0.111 |
| $\tau_{1.2}$ | −0.908 | 0.053 | −0.843 | 0.052 |
| $\tau_{1.3}$ | 0.232 | 0.046 | 0.214 | 0.042 |
| $\tau_{1.4}$ | 1.539 | 0.072 | 1.500 | 0.080 |
| $\tau_{2.1}$ | −2.557 | 0.173 | −2.700 | 0.230 |
| $\tau_{2.2}$ | −1.507 | 0.070 | −1.473 | 0.074 |
| $\tau_{2.3}$ | 0.086 | 0.046 | 0.084 | 0.043 |
| $\tau_{2.4}$ | 1.756 | 0.083 | 1.753 | 0.091 |
| $\tau_{3.1}$ | −2.032 | 0.103 | −2.071 | 0.115 |
| $\tau_{3.2}$ | −1.251 | 0.061 | −1.224 | 0.061 |
| $\tau_{3.3}$ | −0.276 | 0.046 | −0.259 | 0.044 |
| $\tau_{3.4}$ | 1.112 | 0.057 | 1.078 | 0.058 |
| $\tau_{4.1}$ | −2.087 | 0.108 | −2.136 | 0.129 |
| $\tau_{4.2}$ | −1.403 | 0.066 | −1.368 | 0.068 |
| $\tau_{4.3}$ | −0.371 | 0.047 | −0.342 | 0.044 |
| $\tau_{4.4}$ | 1.131 | 0.058 | 1.095 | 0.058 |
| $\tau_{5.1}$ | −2.116 | 0.111 | −2.173 | 0.136 |
| $\tau_{5.2}$ | −1.034 | 0.056 | −0.983 | 0.055 |
| $\tau_{5.3}$ | 0.308 | 0.046 | 0.283 | 0.043 |
| $\tau_{5.4}$ | 1.630 | 0.076 | 1.608 | 0.084 |
| $\tau_{6.1}$ | −0.749 | 0.050 | −0.719 | 0.049 |
| $\tau_{6.2}$ | 0.333 | 0.046 | 0.296 | 0.044 |
| $\tau_{6.3}$ | 1.273 | 0.062 | 1.213 | 0.064 |
| $\tau_{6.4}$ | 2.557 | 0.173 | 2.730 | 0.232 |
| $\tau_{7.1}$ | −1.123 | 0.058 | −1.089 | 0.057 |
| $\tau_{7.2}$ | −0.138 | 0.046 | −0.137 | 0.043 |
| $\tau_{7.3}$ | 0.867 | 0.052 | 0.830 | 0.051 |
| $\tau_{7.4}$ | 1.875 | 0.091 | 1.896 | 0.102 |

*Table F.4.* Threshold parameter and standard error estimates for DBIQ-*Physical contact.* $n = 761$.

| | FA-poly | | IRT-grm | |
|---|---|---|---|---|
| | $\hat{\tau}$ | $\hat{se}(\hat{\tau})$ | $\hat{\tau}$ | $\hat{se}(\hat{\tau})$ |
| $\tau_{1.1}$ | −2.260 | 0.127 | −2.329 | 0.154 |
| $\tau_{1.2}$ | −1.642 | 0.077 | −1.621 | 0.081 |
| $\tau_{1.3}$ | −0.527 | 0.048 | −0.503 | 0.046 |
| $\tau_{1.4}$ | 0.833 | 0.052 | 0.796 | 0.050 |
| $\tau_{2.1}$ | −2.149 | 0.114 | −2.182 | 0.129 |
| $\tau_{2.2}$ | −1.169 | 0.059 | −1.133 | 0.059 |
| $\tau_{2.3}$ | −0.081 | 0.046 | −0.085 | 0.044 |
| $\tau_{2.4}$ | 1.412 | 0.066 | 1.387 | 0.068 |
| $\tau_{3.1}$ | −2.262 | 0.127 | −2.318 | 0.146 |
| $\tau_{3.2}$ | −1.561 | 0.073 | −1.570 | 0.074 |
| $\tau_{3.3}$ | −0.720 | 0.050 | −0.710 | 0.048 |
| $\tau_{3.4}$ | 0.377 | 0.047 | 0.366 | 0.045 |
| $\tau_{4.1}$ | −2.259 | 0.127 | −2.332 | 0.150 |
| $\tau_{4.2}$ | −1.616 | 0.076 | −1.610 | 0.079 |
| $\tau_{4.3}$ | −0.702 | 0.050 | −0.674 | 0.048 |
| $\tau_{4.4}$ | 0.719 | 0.050 | 0.689 | 0.049 |
| $\tau_{5.1}$ | −2.558 | 0.173 | −2.690 | 0.222 |
| $\tau_{5.2}$ | −1.644 | 0.077 | −1.636 | 0.081 |
| $\tau_{5.3}$ | −0.845 | 0.052 | −0.817 | 0.050 |
| $\tau_{5.4}$ | 0.249 | 0.046 | 0.242 | 0.044 |
| $\tau_{6.1}$ | −1.594 | 0.074 | −1.588 | 0.078 |
| $\tau_{6.2}$ | −0.762 | 0.051 | −0.748 | 0.049 |
| $\tau_{6.3}$ | 0.010 | 0.046 | −0.008 | 0.043 |
| $\tau_{6.4}$ | 1.195 | 0.060 | 1.151 | 0.060 |

## F.2 Additional Results for RASJS

### F.2.1 Threshold Parameter Estimation Results

*Table F.5.* Threshold parameter and standard error estimates for RASJS; $n = 1345$

| | FA-poly | | IRT-grm | |
|---|---|---|---|---|
| | $\hat{\tau}$ | $\hat{se}(\hat{\tau})$ | $\hat{\tau}$ | $\hat{se}(\hat{\tau})$ |
| $\tau_{1.1}$ | −1.215 | 0.045 | −1.177 | 0.046 |
| $\tau_{1.2}$ | −0.335 | 0.035 | −0.284 | 0.032 |
| $\tau_{1.3}$ | 0.109 | 0.034 | 0.137 | 0.032 |
| $\tau_{1.4}$ | 0.633 | 0.037 | 0.612 | 0.035 |
| $\tau_{2.1}$ | −0.747 | 0.038 | −0.688 | 0.036 |
| $\tau_{2.2}$ | −0.185 | 0.034 | −0.142 | 0.032 |
| $\tau_{2.3}$ | 0.284 | 0.035 | 0.289 | 0.032 |
| $\tau_{2.4}$ | 0.813 | 0.039 | 0.771 | 0.037 |
| $\tau_{3.1}$ | −2.615 | 0.139 | −2.843 | 0.198 |
| $\tau_{3.2}$ | −2.084 | 0.081 | −2.154 | 0.101 |
| $\tau_{3.3}$ | −1.708 | 0.060 | −1.711 | 0.067 |
| $\tau_{3.4}$ | −1.222 | 0.045 | −1.184 | 0.047 |
| $\tau_{4.1}$ | −1.207 | 0.045 | −1.167 | 0.046 |
| $\tau_{4.2}$ | −0.725 | 0.038 | −0.662 | 0.036 |
| $\tau_{4.3}$ | −0.179 | 0.034 | −0.132 | 0.032 |
| $\tau_{4.4}$ | 0.389 | 0.035 | 0.383 | 0.033 |
| $\tau_{5.1}$ | −0.636 | 0.037 | −0.584 | 0.035 |
| $\tau_{5.2}$ | −0.139 | 0.034 | −0.112 | 0.032 |
| $\tau_{5.3}$ | 0.248 | 0.035 | 0.246 | 0.031 |
| $\tau_{5.4}$ | 0.737 | 0.038 | 0.700 | 0.036 |
| $\tau_{6.1}$ | 0.061 | 0.034 | 0.049 | 0.032 |
| $\tau_{6.2}$ | 0.899 | 0.040 | 0.844 | 0.038 |
| $\tau_{6.3}$ | 1.822 | 0.065 | 1.812 | 0.075 |
| $\tau_{6.4}$ | 2.237 | 0.093 | 2.304 | 0.120 |
| $\tau_{7.1}$ | 0.980 | 0.041 | 0.929 | 0.039 |
| $\tau_{7.2}$ | 1.632 | 0.057 | 1.580 | 0.062 |
| $\tau_{7.3}$ | 2.194 | 0.089 | 2.219 | 0.114 |
| $\tau_{7.4}$ | 2.516 | 0.124 | 2.619 | 0.170 |
| $\tau_{8.1}$ | 0.818 | 0.039 | 0.774 | 0.037 |
| $\tau_{8.2}$ | 1.661 | 0.058 | 1.625 | 0.065 |
| $\tau_{8.3}$ | 2.261 | 0.095 | 2.319 | 0.127 |
| $\tau_{8.4}$ | 2.676 | 0.150 | 2.863 | 0.225 |
| $\tau_{9.1}$ | 0.742 | 0.038 | 0.722 | 0.036 |
| $\tau_{9.2}$ | 1.511 | 0.053 | 1.458 | 0.052 |
| $\tau_{9.3}$ | 2.285 | 0.098 | 2.279 | 0.111 |
| $\tau_{9.4}$ | 2.614 | 0.139 | 2.668 | 0.173 |
| $\tau_{10.1}$ | 0.580 | 0.036 | 0.541 | 0.035 |
| $\tau_{10.2}$ | 1.598 | 0.056 | 1.557 | 0.062 |
| $\tau_{10.3}$ | 2.435 | 0.114 | 2.572 | 0.165 |
| $\tau_{10.4}$ | 2.844 | 0.184 | 3.155 | 0.291 |
| $\tau_{11.1}$ | 0.913 | 0.040 | 0.909 | 0.037 |
| $\tau_{11.2}$ | 1.654 | 0.058 | 1.612 | 0.055 |
| $\tau_{11.3}$ | 2.369 | 0.106 | 2.266 | 0.113 |
| $\tau_{11.4}$ | 2.615 | 0.139 | 2.482 | 0.144 |
| $\tau_{12.1}$ | 0.820 | 0.039 | 0.821 | 0.036 |
| $\tau_{12.2}$ | 1.534 | 0.054 | 1.498 | 0.051 |
| $\tau_{12.3}$ | 2.173 | 0.088 | 2.102 | 0.094 |
| $\tau_{12.4}$ | 2.752 | 0.164 | 2.681 | 0.182 |
| $\tau_{13.1}$ | 0.977 | 0.041 | 0.962 | 0.039 |
| $\tau_{13.2}$ | 1.661 | 0.058 | 1.622 | 0.056 |
| $\tau_{13.3}$ | 2.237 | 0.093 | 2.205 | 0.103 |
| $\tau_{13.4}$ | 2.435 | 0.114 | 2.413 | 0.129 |
| $\tau_{14.1}$ | 0.255 | 0.035 | 0.273 | 0.033 |
| $\tau_{14.2}$ | 0.787 | 0.038 | 0.794 | 0.036 |
| $\tau_{14.3}$ | 1.346 | 0.048 | 1.321 | 0.046 |
| $\tau_{14.4}$ | 1.767 | 0.063 | 1.728 | 0.063 |
| $\tau_{15.1}$ | 0.487 | 0.036 | 0.505 | 0.034 |
| $\tau_{15.2}$ | 1.095 | 0.043 | 1.087 | 0.040 |
| $\tau_{15.3}$ | 1.585 | 0.055 | 1.544 | 0.052 |
| $\tau_{15.4}$ | 2.194 | 0.089 | 2.129 | 0.090 |

### F.2.2 Model Fit Results

*Table F.6.* Model fit results for RASJS-*Reactive*. $n = 1345$.

| Fit statistic | FA-lin | FA-poly | IRT-grm |
|---|---|---|---|
| $\chi^2_{mplus}$ | 16.581 | 8.550 | 2758.972 |
| df | 5 | 5 | 3097 |
| RMSEA | 0.041 | 0.023 | |
| CFI | 0.993 | 0.999 | |
| TLI | 0.986 | 0.998 | |
| $\chi^2_{YB}$ | 16.223 | | |
| df | 5 | | |
| RMSEA | 0.041 | | |
| SRMR | 0.017 | 0.012 | 0.017 |

*Table F.7.* Model fit results for RASJS-*Anxious*. $n = 1345$.

| Fit statistic | FA-lin | FA-poly | IRT-grm |
|---|---|---|---|
| $\chi^2_{mplus}$ | 191.647 | 118.154 | 1648.354 |
| df | 5 | 5 | 3091 |
| RMSEA | 0.167 | 0.130 | |
| CFI | 0.931 | 0.983 | |
| TLI | 0.862 | 0.966 | |
| $\chi^2_{YB}$ | 47.940 | | |
| df | 5 | | |
| RMSEA | 0.080 | | |
| SRMR | 0.051 | 0.043 | 0.048 |

*Table F.8.* Model fit results for RASJS-*Possessive*. $n = 1345$.

| Fit statistic | FA-lin | FA-poly | IRT-grm |
|---|---|---|---|
| $\chi^2_{mplus}$ | 269.457 | 200.787 | 3337.975 |
| df | 5 | 5 | 3091 |
| RMSEA | 0.198 | 0.171 | |
| CFI | 0.873 | 0.961 | |
| TLI | 0.747 | 0.922 | |
| $\chi^2_{YB}$ | 61.804 | | |
| df | 5 | | |
| RMSEA | 0.092 | | |
| SRMR | 0.066 | 0.075 | 0.104 |

# F.3 Additional Results for INCS-*s*

## F.3.1 Threshold Parameter Estimation Results

*Table F.9.* Threshold parameter and standard error estimates for INCS-*s*. $n = 255$.

| | FA-poly | | IRT-grm | |
|---|---|---|---|---|
| | $\hat{\tau}$ | $\hat{se}(\hat{\tau})$ | $\hat{\tau}$ | $\hat{se}(\hat{\tau})$ |
| $\tau_{1.1}$ | −1.674 | 0.135 | −1.673 | 0.147 |
| $\tau_{1.2}$ | −1.148 | 0.101 | −1.118 | 0.101 |
| $\tau_{1.3}$ | −0.399 | 0.081 | −0.378 | 0.077 |
| $\tau_{1.4}$ | −0.015 | 0.078 | −0.013 | 0.075 |
| $\tau_{2.1}$ | −0.684 | 0.086 | −0.664 | 0.083 |
| $\tau_{2.2}$ | −0.530 | 0.083 | −0.516 | 0.079 |
| $\tau_{2.3}$ | −0.431 | 0.081 | −0.420 | 0.078 |
| $\tau_{2.4}$ | −0.064 | 0.079 | −0.065 | 0.075 |
| $\tau_{3.1}$ | −0.485 | 0.082 | −0.470 | 0.080 |
| $\tau_{3.2}$ | −0.253 | 0.079 | −0.241 | 0.076 |
| $\tau_{3.3}$ | −0.025 | 0.078 | −0.016 | 0.075 |
| $\tau_{3.4}$ | 0.474 | 0.082 | 0.467 | 0.079 |
| $\tau_{4.1}$ | −0.899 | 0.091 | −0.888 | 0.090 |
| $\tau_{4.2}$ | −0.325 | 0.080 | −0.309 | 0.076 |
| $\tau_{4.3}$ | 0.243 | 0.079 | 0.256 | 0.075 |
| $\tau_{4.4}$ | 0.960 | 0.093 | 0.938 | 0.091 |
| $\tau_{5.1}$ | 0.015 | 0.078 | 0.013 | 0.076 |
| $\tau_{5.2}$ | 0.103 | 0.079 | 0.101 | 0.075 |
| $\tau_{5.3}$ | 0.335 | 0.080 | 0.333 | 0.077 |
| $\tau_{5.4}$ | 0.734 | 0.087 | 0.723 | 0.084 |
| $\tau_{6.1}$ | 0.420 | 0.081 | 0.374 | 0.078 |
| $\tau_{6.2}$ | 0.599 | 0.084 | 0.540 | 0.080 |
| $\tau_{6.3}$ | 0.842 | 0.089 | 0.775 | 0.087 |
| $\tau_{6.4}$ | 1.148 | 0.101 | 1.089 | 0.100 |
| $\tau_{7.1}$ | 1.207 | 0.103 | 1.160 | 0.106 |
| $\tau_{7.2}$ | 1.501 | 0.121 | 1.463 | 0.125 |
| $\tau_{7.3}$ | 1.635 | 0.132 | 1.609 | 0.141 |
| $\tau_{7.4}$ | 1.808 | 0.149 | 1.801 | 0.165 |

# Samenvatting (Summary in Dutch)

In dit proefschrift worden de verschillen en overeenkomsten tussen factoranalyse (FA) en item-responstheorie (IRT), toegepast op ordinale data, onderzocht met betrekking tot de stabiliteit en sensitiviteit, oftewel robuustheid, van analyseresultaten bij schendingen van verdelingsassumpties. In de inleiding wordt de gangbare wijze waarop FA- en IRT-modellen worden gebruikt voor schaalconstructie en -evaluatie toegelicht. Vervolgens worden de belangrijkste onderzoeksvragen behandeld en de keuze voor de onderzoeksopzet onderbouwd.

In Hoofdstuk 1 wordt na een korte introductie van het begrip latente variabele (LV) een inleiding gegeven in de twee te onderzoeken benaderingen van schaalanalyse: FA en IRT. Vanuit twee onafhankelijke tradities hebben zich deze twee modellen voor schaalanalyse ontwikkeld, die in een aantal opzichten gelijk zijn aan elkaar. Sommige FA- en IRT-modellen zijn zelfs mathematisch equivalent. Vanwege hun verschillende oorsprong worden bij FA en IRT verschillende schattingsmethoden toegepast, waarbij voor FA met name zogenaamde 'limited-information' methoden worden gebruikt en voor IRT daarentegen vooral 'full-information' methoden.

De toepassing van schaalanalyse in de praktijk is onderwerp van Hoofdstuk 2. Er worden 40 wetenschappelijke artikelen besproken, waarin FA- en IRT-modellen zijn gebruikt, uit jaargang 2005 van drie tijdschriften op het gebied van schaalontwikkeling en -evaluatie. Zoals verwacht, wordt FA hierin veel vaker toegepast dan IRT. Onderzoekers verantwoorden hun modelkeuze zelden, waardoor we genoodzaakt zijn te speculeren over hun motieven voor zo'n keuze. De verwachting dat een schaal multidimensioneel is zou een reden kunnen zijn om FA te gebruiken in plaats van IRT. Een ander mogelijk motief voor de voorkeur van FA boven IRT is het gebrek aan toegankelijke software voor (multidimensionele) IRT. Modelassumpties, zoals aannames over de verdeling van de data, worden vaak niet onderzocht of worden in ieder geval niet vaak gerapporteerd. Dit is zorgelijk, aangezien er vaak een lineair FA-model wordt toegepast op ordinale gegevens, hetgeen alleen tot acceptabele resultaten leidt indien de items bij benadering normaal verdeeld zijn, zoals uit eerder onderzoek is gebleken. Deze empirische bevindingen uit de praktijk geven aanleiding tot vragen over de consequenties van het gebruik van schaalmodellen wanneer hun assumpties

zijn geschonden en vormen daarmee een belangrijke praktische motivatie voor dit promotieonderzoek.

In Hoofdstuk 3 wordt een overzicht gegeven van eerder simulatieonderzoek over FA van ordinale data, twee-parametrische IRT-modellen of beide. Die studies worden samengevat en chronologisch geordend weergegeven in twee overzichtelijke tabellen (zie pp. 67 en 68). Op basis van de besproken literatuur worden algemene verwachtingen opgesteld over de in eigen simulatieonderzoek te beantwoorden robuustheidsvragen en worden de verklarende factoren van ons onderzoeksontwerp, inclusief de gekozen niveaus daarvan, verantwoord: (a) LV-verdeling, (b) item-responsverdeling, (c) schaalsterkte en (d) steekproefomvang. De dataconfiguraties, die volgen uit het combineren van deze factoren, zullen worden geanalyseerd met vier schaalmodellen: FA van de steekproefcovariantiematrix met een meest aannemelijke schattingsmethode (FA-lin-ML), FA van de geschatte polychorische-correlatiematrix met een 'mean-and-variance adjusted weighted least squares' schattingsmethode (FA-poly-WLSMV), het 'graded response' model met een robuuste meest aannemelijke schattingsmethode (IRT-grm-MLR) en het niet-parametrische Mokken IRT-model (IRT-mok).

De opzet van het simulatieonderzoek wordt nader uitgewerkt in Hoofdstuk 4, aan de hand van het proces van datageneratie, de uitkomstvariabelen en de toe te passen performance criteria bij de evaluatie van de simulatieresultaten. Bovendien worden onze verwachtingen gespecificeerd met betrekking tot de uitkomstvariabelen, die in grote mate gebaseerd zijn op de besproken literatuur in Hoofdstuk 3. Het onderzoeksontwerp wordt daarbij opgedeeld in tweeën: normale dataconfiguraties (met bij benadering normaal verdeelde, maar categorische items die laden op een normaal verdeelde LV) worden behandeld in Hoofdstuk 5; niet-normale dataconfiguraties worden besproken in Hoofdstuk 6.

We verwachten slechts kleine verschillen in resultaten tussen de vier modellen voor de normale datacondities, los van een kleine doch consistente negatieve onzuiverheid voor FA-lin ladingparameterschatters, bekend uit eerder onderzoek. Bij afwijkingen van normaliteit van LV- en/of itemverdelingen verwachten we grotere verschillen in resultaten tussen de modellen. Voor FA-lin wordt de grootste afwijking verwacht, met een aanzienlijke onderschatting van parameters, behalve in het geval van gelijkelijk over de latente schaal verdeelde drempelwaarden, oftewel wanneer de LV en de items eenzelfde scheefheid vertonen. We verwachten dat FA-poly parameterschatters meer zullen worden gehinderd door niet-normaliteit van de LV-verdeling dan door niet-normaliteit van de itemverdelingen. Onder niet-normaliteit worden van IRT-grm de beste resultaten verwacht in vergelijking met de overige parametrische modellen. Vanwege zijn niet-parametrische karakter, verwachten we niet dat IRT-mok duidelijk beïnvloed wordt door niet-normaliteit.

De resultaten van het toepassen van de vier schaalmodellen op gegenereerde normale data worden besproken in Hoofdstuk 5. De gegenereerde data bestaan uit unidimensionele schalen van 12 items met vijf antwoordcategorieën. Items zijn ofwel alle sterk geassocieerd met de LV ofwel van gemêleerde samenstelling met vier sterk-,

vier matig- en vier zwak-geassocieerde items. De steekproefomvang is klein ($n = 200$) of middelgroot ($n = 600$).

In Tabel 5.11 (p. 139) worden de resultaten van de parametrische modellen voor de normale dataconfiguraties samengevat met verwijzing naar de in Hoofdstuk 4 gepresenteerde verwachtingen. De meeste verwachtingen worden ondersteund. Twee uitzonderingen betreffen de hypothesen over FA-lin schatters van standaardfouten en LV scores, die nauwkeuriger zijn dan verwacht. FA-poly en IRT-grm doen het, in overeenstemming met de verwachtingen, goed voor alle uitkomstvariabelen.

Voor het niet-parametrische IRT-mok-model vinden we dat schaalbaarheidsco-efficiënt Loevinger's $H$ consistent in kleine mate wordt overschat met ca. 5%, aflopend naar nul voor zeer grote steekproeven. Vermoedelijk wordt deze onzuiverheid veroorzaakt door het feit dat in de datageneratie alle populatiewaarden van de itemgemiddelden gelijk zijn gekozen, hetgeen de berekening van $H$ bemoeilijkt aangezien die gebaseerd is op itemordening in een steekproef. De schatters van de standaardfouten voor $H$ zijn zuiver.

In Hoofdstuk 6 wordt de simulatiestudie aangaande niet-normaliteit beschreven. Dataconfiguraties verschillen in de itemverdelingen (normaal, scheef of bimodaal), de LV-verdeling (normaal of scheef-normaal) en steekproefomvang (klein of middelgroot), waarbij de item-LV-associaties constant sterk worden gehouden.

De resultaten voor de parametrische modellen worden samengevat in Tabel 6.11 (p. 188), opnieuw met verwijzing naar de in Hoofdstuk 4 gepresenteerde verwachtingen. Zoals duidelijk wordt uit het linkerdeel van deze tabel, geven FA-poly en IRT-grm goede resultaten voor elke uitkomstvariabele bij een normale LV, ongeacht de itemverdelingen, en doet FA-lin het aanzienlijk slechter. Bij een scheef-normale LV verslechteren de resultaten van alle parametrische modellen, met name wanneer de itemvariabelen ook scheef verdeeld zijn. In dit geval vertoont IRT-grm de beste resultaten met een nauwkeuriger schatting van parameters en standaardfouten dan FA-poly.

IRT-mok is robuust tegen niet-normale LV- of itemverdelingen. Het is zelfs zo dat schalen met items van variërende vorm leiden tot hogere $H$-waarden, die bovendien nauwkeuriger geschat worden dan bij homogene schalen. Schatters van standaardfouten zijn zuiver onder alle condities.

De resultaten, zoals gepresenteerd in Hoofdstuk 5 en 6, geven duidelijk aan dat LV-scores op basis van geschatte modelparameters meer informatie geven over de ware LV-score van een respondent dan ongewogen somscores, vooral bij incongruente LV- en itemverdelingen (bijv. scheve items die laden op een normale LV) of schalen met items die verschillen in lading.

In Hoofdstuk 7 keren we terug naar de toepassing van schaalanalyse in de praktijk. We bespreken een aantal toepassingen van de te onderzoeken schaalmodellen, waarbij we een gedetailleerde illustratie geven van het gebruik van FA-lin, FA-poly, IRT-grm en IRT-mok in drie empirische settings. Door zorgvuldig de itemverdelingen te onderzoeken en de resultaten van de schaalmodellen (inclusief geschatte LV-scores) te vergelijken kunnen we (voorzichtige) conclusies trekken over de LV-verdeling

op basis van de bevindingen van ons simulatieonderzoek. Hoewel de populatiewaarden van de parameters uiteraard onbekend zijn in de empirische praktijk, blijkt het meestal mogelijk de resultaten te relateren aan de bevindingen uit de simulatiestudie, waarmee de interpretatie van de schattingsresultaten kan worden verbeterd.

In het laatste hoofdstuk worden de resultaten van ons onderzoek samengevat. Daarnaast geven we praktische richtlijnen voor het toepassen van de schaalmodellen. We adviseren onderzoekers een voldoende grote steekproefomvang te kiezen, gebruik te maken van hun inhoudelijke kennis over de schaal, zorgvuldig de verdeling van de steekproefgegevens te bestuderen, op basis hiervan een schaalmodel te selecteren, de passing van het model na te gaan en LV-scores op basis van geschatte modelparameters te gebruiken voor vervolganalyses. Het hoofdstuk, en daarmee het proefschrift, wordt afgesloten met een verkenning van de implicaties van dit onderzoek en aspecten van bevindingen die nader onderzoek verdienen.

# Dankwoord (Acknowledgements)

Het heeft even geduurd, maar nu ligt het er.

Dit proefschrift is totstandgekomen in nauwe samenwerking met mijn twee primaire begeleiders, Anne Boomsma en Marijtje van Duijn. Ik bedank hen hartelijk voor hun bijdrage aan dit project. Anne, je enthousiasme tijdens je college over waarschijnlijkheidsrekening was zeer aanstekelijk. Het heeft er mede toe geleid dat ik ben gaan werken aan dit project. Ik ben je begeleiding door de jaren heen steeds meer gaan waarderen en zal onze besprekingen met anecdotes over bijvoorbeeld verkeersboetes en kappers, soms onderbroken doordat je aandacht werd getrokken door een vogeltje dat te zien was door die enorme ramen aan de Grote Rozenstraat, missen. Je grenzeloze nauwkeurigheid bij het becommentariëren van stukken is ongekend en is de kwaliteit van het proefschrift zeer ten goede gekomen. Hoewel ik weet dat het eindresultaat onmogelijk geheel naar je zin kan zijn — er staan vele figuren en tabellen niet precies waar jij ze wilt hebben — hoop ik dat je tevreden bent over "je laatste aio".

Marijtje, ik waardeer je meedenken en hulpvaardigheid enorm. De samenvattende tabellen met daarin nagenoeg alle simulatieresultaten, die ik zo handig en overzichtelijk vind, zijn er dankzij jouw ingeving. De deur van jouw kamer stond altijd open. Dit ging zo ver dat ik in de laatste fase van het project zelfs een plekje op je kamer heb gekregen, waar ik op jouw vrije woensdag in alle rust aan het proefschrift kon werken. Als je wel aanwezig was op zulke woensdagen, was je gelukkig veel in bespreking, want als we samen op je kamer zaten, was het moeilijk om niet uren te zitten praten over zaken als kinderen, moeders, werkende moeders en nog veel meer.

Uiteraard bedank ik ook mijn promotor, Tom Snijders, voor zijn bijdrage aan dit proefschrift. Tom, je bent tijdens het leeuwendeel van het project een "promotor op afstand" geweest en vertrouwde mijn begeleiding terecht toe aan Anne en Marijtje. In de laatste fase ben je aangeschoven bij de besprekingen waarin we alle hoofdstukken een voor een doornamen. Ik heb je frisse blik hierbij zeer gewaardeerd. (Bovendien heb ik iets kunnen leren over de bijzondere herkomst van de Box-Cox transformatie.)

Naast mijn begeleiders wil ik de collega's aan de faculteit bedanken voor het creëren van een prettige werkomgeving. Door de jaren heen heb ik vele werkplekken gehad en elke gang bracht een andere groep collega's met zich mee, waartussen ik me iedere keer opnieuw thuis ging voelen.

Ook bedank ik mijn stief-, schoon- en "gewone" ouders voor hun onaflatende steun tijdens dit project. Het zal even wennen zijn niet langer de vraag te hoeven stellen: *"Hoe is het met je proefschrift?"*. Hanneke, veel dank voor het maken van de prachtige omslagillustratie.

Dan tot slot nog een dankbetuiging aan mijn echtgenoot. Mick, mijn steun en toeverlaat, dank je wel voor het faciliteren van mijn tijdrovende "hobby", op vele vlakken: van IT-ondersteuning tot omslagontwerp tot zorgen voor de kinderen.

Bram, Marjolein, het is nu *echt* klaar!